Quantum Mechanical Methods for Structure-Based Drug Design

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Ting Zhou
aus
China

Promotionskomitee

Prof. Dr. Amedeo Caflisch
Prof. Dr. Kim Baldridge

Zürich 2010

Inhaltsverzeichnis

Abstract

Quantum mechanical (QM) methods are becoming popular in computational drug design and development mainly because high accuracy is required to estimate (relative) binding affinities. For low- to medium-throughput in silico screening, (e.g., scoring and prioritizing a series of inhibitors sharing the same molecular scaffold) efficient approximations have been developed in the past decade, like linear scaling QM in which the computation time scales almost linearly with the number of basis functions. The first chapter of this thesis is a review of QM methods for drug design.

In the second chapter, LIECE, which is short for linear interaction energy model with continuum electrostatic solvation, is further improved by using a linear scaling semiempirical QM method. The new method is termed QMLIECE. Then QMLIECE is tested on three enzyme/inhibitor systems: the West Nile virus NS3 protease (a serine protease), the HIV-1 protease (an aspartic protease), and the human cyclindependent kinase 2. After that the necessity of QM method for predicting binding free energy is discussed.

In the third chapter, the QM probe method is suggested as a filter in high throughput virtual screening. To speed up the calculation, the QM probe method approximates the ATP-binding site of the tyrosine kinase erythropoietin producing human hepatocellular carcinoma receptor B4 (EphB4) by 6 types of small molecules (probes) with particular orientations, and uses a semi-empirical QM Hamiltonian to calculate interaction energies between a compound and probes. These interaction energies are further used for filtering multimillion poses generated by high throughput docking. A single-digit micromolar inhibitor of EphB4 with a relatively good selectivity profile is identified upon experimental tests of only 23 molecules.

To elucidate the technology of high throughput docking used in the third chapter, in the forth chapter, a distributed virtual screening data management system (DVSDMS) is introduced, in which the data handling and the distribution of jobs are realized by the open-source structured query language database software MySQL. In DVSDMS, the data management is separated from docking and ranking process. In both benchmark and production, DVSDMS performed efficiently with a limited effort of programming and a trivial investment of software and hardware.

In the last chapter, the original ultrafast shape recognition (USR) method is complemented with an optical isomerism descriptor (USR:OptIso). Then the USR:OptIso is tested on discriminating mirror images of three kinase inhibitors and 15 types of isomers, some of which can not be distinguished by the original USR, which uses only the atomic distances. Finally, both similarity scores calculated by the original USR and USR:OptIso are extensively compared with ROCS shape Tanimoto, which is based on Gaussian molecular volume overlap.

Zusammenfassung

Quantenmechanische (QM) Methoden erfreuen sich im computergestützen Wirkstoffdesign und der Entwicklung wachsender Beliebtheit, da eine hohe Genauigkeit für die Berechnung von Bindungsaffinitäten notwendig ist. Für das Screening einer kleinen und mittleren Anzahl an Molekülen wurden im letzten Jahrzehnt effiziente Näherungen entwickelt.

In dieser Arbeit wird LIECE, was kurz ist für lineares Interaktions-Energiemodell mit kontinuierlicher elektrostatischer Solvatation, weiterentwickelt durch Verwendung einer linear skalierenden, semiempirischen QM Methode. Das neue Verfahren wird QMLIECE genannt. QMLIECE wird an drei Enzym/Inhibitor-Komplexen getestet. Anschliessend wird die Notwendigkeit der QM Methode für die Vorhersage der freien Bindungsenergien diskutiert.

Im nächsten Abschnitt der Arbeit wird die QM Probemethode als Filter für virtuelles Hochdurchsatz-Screening vorgeschlagen, für das ein neues virtuelles Screening-Datenverarbeitungssystem angewendet wird. Ein einstellig mikromolarer Inhibitor von EphB4 mit relativ gutem Selektivitätsprofil wird identifiziert durch experimentelle Untersuchung von nur 23 Molekülen.

Chapter 1

Quantum Mechanical Methods for Drug Design

Introduction

Accurate models for computing the binding free energy between small molecules and proteins are needed for drug discovery and design.[1] The increasing popularity of quantum mechanical methods in computer-aided drug design (CADD) is not just a consequence of ever growing computing power but is also due to the first principle nature of QM, which should provide the highest accuracy.[2, 3] Because of their first principle nature, both the time-consuming ab initio methods[4] and fast semi-empirical approaches[5, 6] do not suffer from the limitation inherent to the ball and spring description and the fixed-charge approximation used in the force fields (FFs). In a recent review, it has been suggested that the application of QM methods in all phases of CADD is likely to become reality.[3] At the same time, interest for QM in CADD has spurred further methodological development of QM methods and in particular QM approaches for docking, scoring, improvement of known lead compounds, and unraveling the reaction mechanism. As an example, QM calculations were performed to investigate significant differences in binding affinities upon modification of a $-CH_2-$ linker into a carbonyl.[7]

In the following sections, we will classify the QM methods into two broad classes according to their functionalities: the first class includes the methods used for quantifying energies and optimizing structures, while the second contains the techniques employed for calculating molecular properties. The methods in the first class are the conventional and straightforward applications of QM. They can be directly exploited for interpreting the reactivities of biologically active molecules, which are always accompanied by the transfer of energy and transformation of molecular structures. However, the currently available computating power is not enough for the direct ab initio QM calculations of macromolecules with accuracy similar to that of in vitro experiments. Therefore this section will unavoidably give the prominence to the acceleration of QM methods for macromolecules, including linear scaling algorithms and hybrid quantum-mechanics/molecular-mechanics (QM/MM). Apart from the methods for calculating the energies and optimizing structures of biomolecules, we also illustrate two typical applications of QM relevant to structure-based drug design: the analysis of the protonation states of titratable side chains[8, 9] and the evaluation, with high structural and energetic

accuracy, of cation-π and π-π interactions that are beyond the limits of classical FF methods. The methods in the second class are mainly used for calculating specific properties of molecules, such as partial charges, bond strength, and torsion angles which can be applied in the parameterization of FFs, and other descriptors that can be used in building quantitative structure-activity relationship (QSAR) models or quantitative structure-property relationship (QSPR) models. We also discuss recent advances in two emerging topics: molecular quantum similarity and variational particle number approach for molecular design.

Using QM to calculate energies and optimize structures

QM is preferable to classical FF based methods for accurate energies and electronic structure calculations,[2, 3] and even for examining the potential energy hypersurface of small molecules.[10] Recently, Butler and coworkers used QM-based methods to describe both the internal energy of the ligand and the solvation effect.[11] Their analysis indicates that two thirds of the bioactive conformations of small-molecule inhibitors lie within 0.5 kcal·mol^{-1} of a local minimum, and conformations with penalties above 2.0 kcal·mol^{-1} are generally attributable to inaccuracies in structure determination. However QM can only be applied to molecular systems of limited size, so that QM has to be simplified to adapt to the available computational power, e.g., apply pure QM to a small subset of atoms and polarizable continuum model to emulate the protein and the solvent,[12] or use accelerated techniques which will be mentioned in the following sections.

Linear scaling QM methods

The computational time of QM ranges from N^3 (semi-empirical) to N^5 (second order Møller-Plesset perturbation theory (MP2) and other post-Hartree-Fock (HF) methods), where N is the number of basis functions.[13] Linear scaling quantum mechanics (LSQM) has been applied extensively for the evaluation of binding enthalpy between small molecules and proteins.[2, 3] In LSQM, the computing time scales with N^2 or even N if the local character of chemical interactions is fully

exploited.[14–18] In the divide-and-conquer (D&C) approach, one of the typical LSQM techniques, a large system is decomposed into many subsystems, and the density matrix of each subsystem is determined separately. Finally contributions of individual subsystems are summed to obtain the total density matrix and energy of the system (1).[15–17, 19] Raha and Merz developed a semi-empirical D&C-based scoring function[18] and studied the ion-mediated ligand binding process. Their study shows that QM is needed for metal-containing system, because the atom types and parameters of metal atoms in most classical FFs are not accurate enough to describe the nature of the interactions between a small molecule and a metal ion in the active site.[20]

We have suggested the use of a semi-empirical D&C strategy as an improvement of the linear interaction energy model with continuum electrostatic solvation (LIECE).[21] The new method QMLIECE was compared to LIECE by application to three enzymes belonging to different classes: the West Nile virus NS3 serine protease (WNV PR) , the aspartic protease of human immunodeficiency virus (HIV-1 PR), and the human cyclin-dependent protein kinase 2 (CDK2).[19] Our results indicate that QMLIECE is superior to LIECE when the inhibitor/protein complexes have highly variable charge–charge interactions, as in the case of 44 peptidic inhibitors of WNV PR, because of the variable polarization effects (2) which are captured only by QMLIECE (3).

Localized molecular orbital (LMO) theory is an LSQM approach in which occupied-virtual interactions involving distant LMOs are neglected, i.e., only density matrix and energies of LMOs that belong to a limited number of atoms need to be calculated.[28] Used as a scoring function in virtual screening, however, LMO theory is still not fast enough for evaluating all poses of small molecules generated by docking. Therefore, Vasilyev and Bliznyuk, performed QM implemented in MOZYME[28] based on LMOs only for the 10 – 100 top binders predicted by simplified scoring functions.[29] By comparing the results with and without solvation they pointed out that although QM was able to provide more accurate enthalpy values, the solvation model needed to be improved. Anikin and coworker developed another linear-scaling semi-empirical algorithm based on LMOs named LocalSCF.[30] The method resolves the SCF task through the finite atomic expansion of weakly nonorthogonal localized molecular orbitals. The inverse overlap matrix arising from the

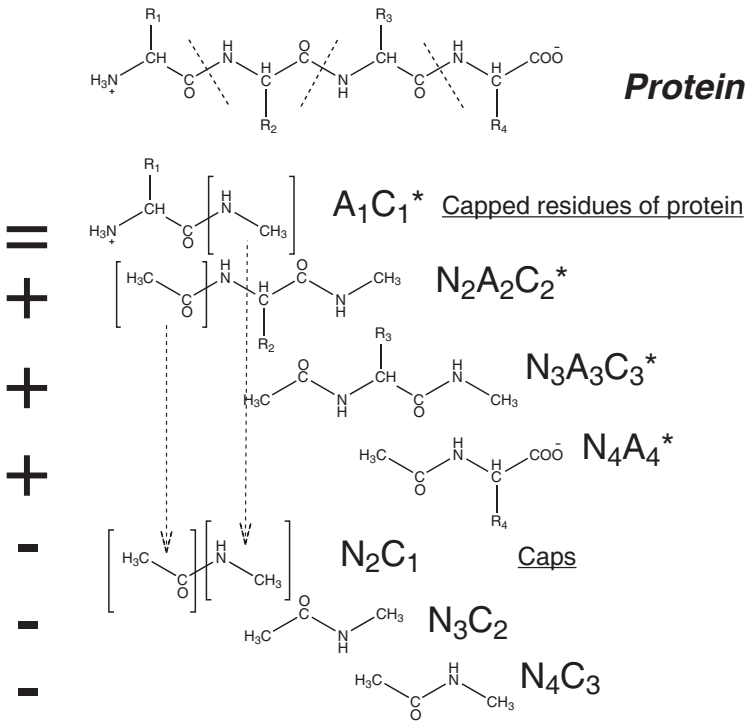


Figure 1: D&C protocol[15] for calculation of QM interaction energy between a protein and a small molecule (ligand). The interaction energy between a protein with m residues and the ligand is decomposed into

$$E_{\text{ligand-protein}} = E_{\text{ligand-A}_1\text{C}_1} + E_{\text{ligand-N}_2\text{A}_2\text{C}_2} + \ldots + E_{\text{ligand-N}_{m-1}\text{A}_{m-1}\text{C}_{m-1}} + E_{\text{ligand-N}_m\text{A}_m} - E_{\text{ligand-N}_2\text{C}_1} - E_{\text{ligand-N}_3\text{C}_2} - \ldots - E_{\text{ligand-N}_m\text{C}_{m-1}}$$

where N_i and C_i are N terminal and C terminal cap, respectively, of residue A_i. The fragments with blue names are protein residues with conjugate caps,[17] while the ones with red names are pure "caps" that have to be subtracted to remove the duplication in energy calculation.[19] This figure is reprinted from Ref.19 with permission of ACS.

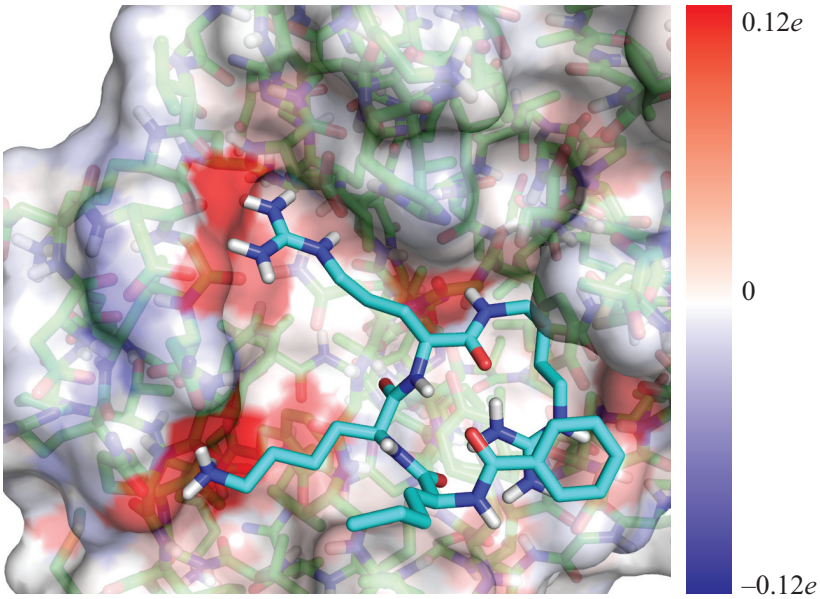


Figure 2: Polarization of protein atoms due to inhibitor binding: WNV PR, whose carbons are in green, in complex with a peptidic inhibitor, whose carbons are in cyan. The polarized charges were calculated by subtracting self-consistent field (SCF) atomic charges before binding from that after binding, using the D&C protocol (see 1).[15] The protein surface was rendered with the blue-white-red spectrum according to polarized charges of atoms. The blue color on the surface denotes atomic partial charges that become more positive upon binding, while red color means more negative atomic charges upon binding, and white color indicates atomic charges which do not change upon binding.

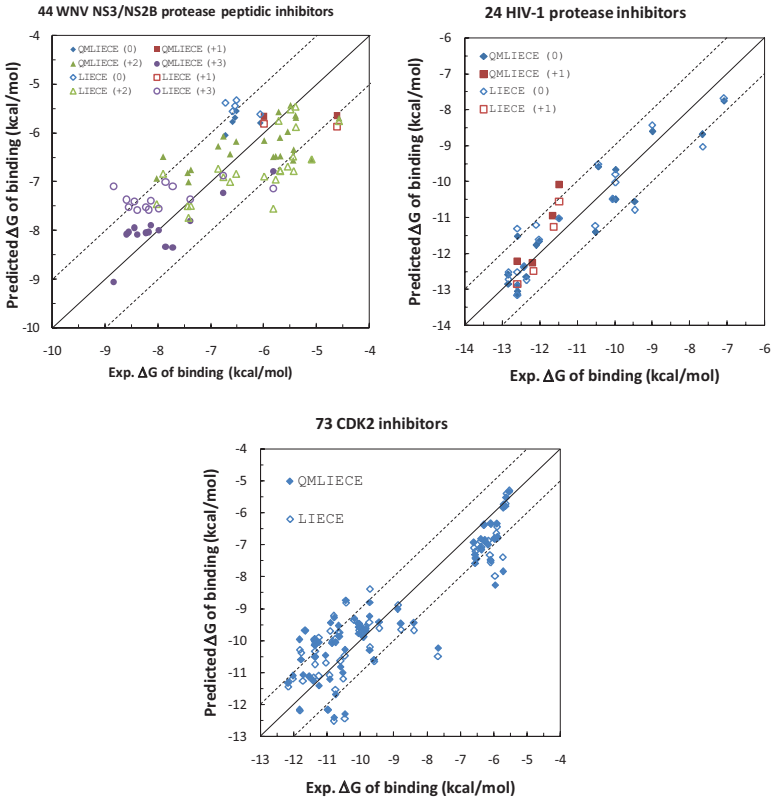


Figure 3: Comparison of the calculated (QMLIECE filled symbols, LIECE empty symbols) versus experimental binding free energies for 44 WNV PR[22–24] (top left), 24 HIV-1 PR[25] (top right), and 73 CDK2[26, 27] (bottom) inhibitors. The experimental data are fitted with two-parameter models for WNV PR, three-parameter models for HIV-1 PR, and two-parameter models for CDK2. Digit in parentheses is the total charge of the inhibitor. This figure is reprinted from Ref.19 with permission of ACS.

nonorthogonality of the localized orbitals is approximated by preserving the first-order perturbation term and applying the second-order correction by means of a penalty function. Furthermore, for very large systems ($>$ 104 atoms) the performance of LocalSCF is more efficient in CPU time and memory consumption than MOZYME with the help of the fast multipole method.[31]

QM/MM

The computational procedures based on QM/MM (see also the review article by J. Gascon at page nnn) combine the strengths of both QM (accuracy) and molecular mechanics (MM) (efficiency) methods, and are widely employed to model chemical reactions and other electronic processes in biomolecular systems.[32–42] QM/MM can be used for preparing the structures of small molecules and proteins, such as optimizing the binding poses obtained from docking,[43] and refining the geometries of enzyme active sites obtained from a harmonically restrained minimization with MM,[44] or X-ray structures.[45] It has been suggested that within the drug discovery process QM/MM is valuable for (1) helping the interpretation of poorly resolved electron density,[45] (2) probing the details of the interactions within enzymes active sites,[46] and (3) investigating the effects of different substituents on the binding mode or in the assessment of alternate scaffolds.[47]

QM/MM is also very useful in describing the process of charge polarization and electron transfer, which is not possible by classical FF methods. Anisimov and coworkers used a QM/MM docking approach based on variational finite localized molecular orbital approximation to speed up conventional QM,[30, 31] which took explicitly into account the effects of charge polarization and intermolecular charge transfer.[48] Gentilucci et al. carried out QM/MM calculation to investigate the binding mode into the M-opioid receptor and the electronic properties of an atypical agonist, the cyclic peptide c[YpwFG] which contains aromatic side chains.[43] In their research the highly favorable dipole-dipole interaction between the protein and the peptide agonist indicates that ligand polarization induced by the protein environment contribute noteworthily to the overall binding energy. Gao and coworkers used docking, molecular dynamics (MD), and QM/MM methods to study the reaction dynamics between pyrimidine nucleoside phosphorylase and a substrate.[49]

Their results show that catalysis involved residues stabilize the uridine in a high-energy conformation by electrostatic interactions and the activation of phosphorolytic catalysis stems from polarization effects. As mentioned above, QM is also appropriate to describe electron transfer. Blumberger and coworkers applied QM/MM to calculate the free energy profile for peptide bond cleavage.[50] Zheng and coworkers developed a QM/MM based approach for in silico screening of transition states of the enzymatic reaction and calculated the activation energy.[51] By this approach they designed a human butyrylcholinesterase mutant with a 2000-fold improved catalytic efficiency for therapeutic use as an exogenous enzyme in humans to treat cocaine overdose and addiction. Wallrapp and coworkers presented docking and QM/MM studies for the electron transfer pathway between cytochrome P-450 camphor and putidaredoxin.[52]

QM/MM is a powerful instrument of parameterization of FFs for the system containing structural motifs not adequately described by empirical FFs, such as diverse drug-like molecules,[53] and metal-containing system.[54] QM/MM can also be used for incorporating polarization effects into a FF, which enables the qualitative improvement in constructing patterns of hydrogen bonds of the docked ligand, water structures and dynamics.[55]

QM Simulation

QM simulation is a useful tool for unraveling the mechanism of reactions.[56] In the drug design field which involves biological macromolecules, QM simulation is often working with the classical MD simulation.[40, 57] To explain the catalytic pathway of metalloenzyme farnesyltransferase (FTase), Ho and coworkers exploited the Car-Parrinello MD[58] version of QM(B3LYP density functional theory (DFT))/MM(Amber FF[59]) dynamics. Their results might be helpful in designing selective inhibitors of FTase, given the proposed mechanism of the FTase reaction and the inhibition by fluorine substituents of farnesyl diphosphate substrate.[60]

QM combined with classical MD is also useful for improving accuracy of interaction energy and sampling of conformational space. Feenstra and coworkers used semi-empirical QM to calculate activation energy barriers, and compare substrate activation barriers at different locations from MD

simulations in the enzyme.[61] Alves and coworkers explained the viral resistance of diketo acids (DKAs) to the integrase of human immunodeficiency virus (HIV-1 IN) N155S mutant by QM/MM MD simulation.[62] Their decomposition analysis of energy terms shows that there is a strong interaction between the Lys159, Lys156, Asn155, and Mg^{2+} cation and the DKA inhibitor with complex electrostatic interactions. QM/MM can be used in free energy perturbation (FEP) method. QM/MM FEP was applied to calculate the relative solvation free energies for a diverse set of small molecules (root mean square deviation (RMSD)) from experimental data $< 1.02\,\mathrm{kJ{\cdot}mol^{-1}}$).[53] Using the same method, the $>$ 2000-fold decrease in the affinity for fructose-1,6-bisphosphatase of an adenosine monophosphate (AMP) analogue (phosphonate 4) compared with AMP was explained by the absence of hydrogen bonds and the loss of the electrostatic interactions,[63] which were well described by the QM method.[64] Similarly Khandelwal and coworkers reported that QM/MM calculated energies for the time averaged structures from MD simulations were able to distinguish subtle differences in binding affinities of only one order of magnitude (4).[65, 66] These methods, however, are very time-consuming, and therefore are not applicable for in silico high throughput screening at present.

Protonation states

The rapid growth of the number of protein structures determined by X-ray crystallography calls for robust methods for determining hydrogen positions, in particular for active site residues in enzymes.[35, 36, 67] Explicit hydrogen atoms are required for most of the structure based drug design methods,[68] e.g., all-atom MM, MD, docking, and electrostatic calculations. A recent study reaffirms that the protonation state in the active site influences the ability of scoring methods to determine the native binding pose.[69] Although other classical methods, e.g., MD[67] and MM/Poisson-Boltzmann (PB) surface area,[70] can be used for determining the position of hydrogen atoms, the prediction of protonation states should be more robust by means of QM because protonation is related to the formation of the covalent bond between the hydrogen and heavy atom.

There are several studies on the determination of protonation states of protease, e.g., β-secretase

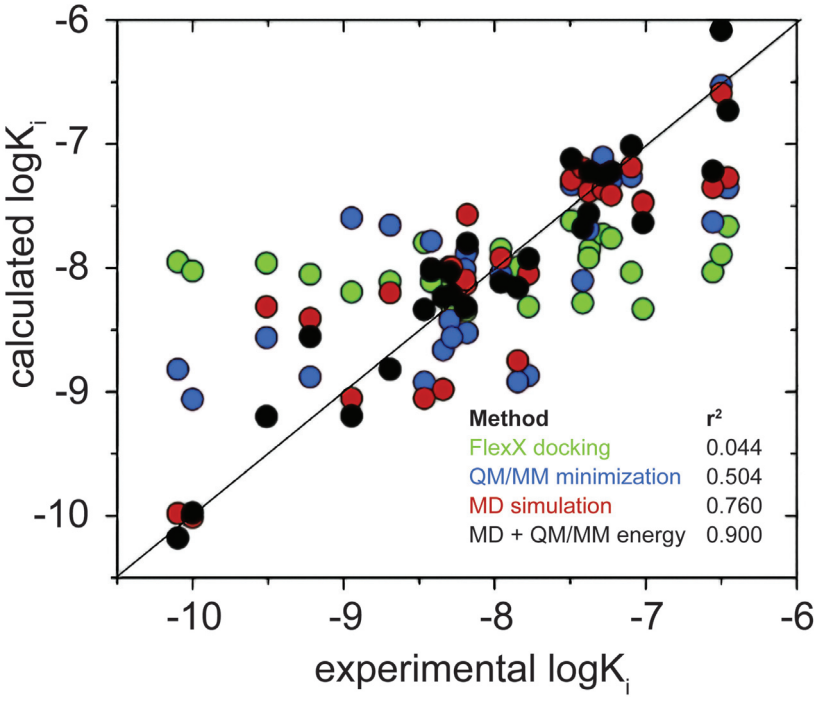


Figure 4: Correlations between experimental and calculated inhibition potencies of hydroxamates vs MMP-9 as obtained by FlexX docking with the zinc-binding-based selection of modes (green), QM/MM minimization (blue), MD simulation with constrained zinc bonds (red), and by QM/MM energy calculations for the time-averaged structures from MD simulation (black).[65, 66] This figure is reprinted from Ref.65 with permission of ACS.

(BACE),[71–74] plasmepsin,[67] and HIV-1 PR.[69, 75] Which of the two aspartates in the catalytic dyad of BACE (Asp32 or Asp228) is protonated is likely to depend on the presence and type of inhibitor.[71, 72, 76] Rajamani and coworkers used a LSQM method and the finite-difference PB method to determine the protonation state and proton location in the presence and absence of an inhibitor.[72] They performed structural optimization in the region surrounding the catalytic dyad. Their calculation favors the monoprotonated state of Asp228 in presence of the hydroxyethylene based inhibitor and di-deprotonated state for the apo enzyme. Yu and coworkers applied the QM/MM to further refine the X-ray structure of BACE, and observed an energetically favored monoprotonated configuration of Asp32 by fitting eight refined structures of BACE and an inhibitor to the observed electron density.[74]

Determination of protonation states of metal-binding sites poses challenges on classical methods.[77, 78] Lin and Lim used a combination of QM and continuum dielectric methods to compute the free energies for deprotonating a Zn-bound imidazole/water in various zinc complexes.[79] They found that the protonation state of the His in the Zn-binding site depends on the solvent accessibility of metal-binding site and Lewis acid ability of the zinc atom. They also suggested that it is critical for the QM region to include not only the metal's first-shell interactions, but also the second shell in QM/MM modeling of metal-binding sites of metalloproteins.[79, 80] A comprehensive review of Kamerlin and coworkers summarizes the progresses in ab initio QM/MM free-energy simulations of electrostatic energies in proteins.[40] Their accelerated QM/MM method, which uses an updated mean charge distribution and a classical reference potential, was benchmarked on the pK_a of titratable side chains. For Asp3 in the bovine pancreatic trypsin inhibitor they obtained the deviation of ~ 1 pK_a unit ($1\,\mathrm{kcal{\cdot}mol^{-1}}$). For Lys102 in T4-lysozyme mutant the deviation was 2.4 pK_a unit ($\sim 3\,\mathrm{kcal{\cdot}mol^{-1}}$). The protonation state of Lys102 may affect the conformation of the protein, since it is deeply buried in the hydrophobic surface. Therefore there is much larger likelihood to attain significant errors in calculation of pK_a of its side chain.[81] Compared to the $7\,\mathrm{kcal{\cdot}mol^{-1}}$ energy difference required for catalysis, an error of $3\,\mathrm{kcal{\cdot}mol^{-1}}$ may be acceptable to determine the main energetic contribution to the reaction.

Cation-π and π-π interactions

Cation-π and π-π stacking interactions play a fundamental role in chemical and biological recognition.[82] Classical FFs sometimes fail to describe these interactions because of the lack of charge delocalization in fixed-charge models or the particular FF parameters. Even HF methods have limitations in capturing π-interactions because of incompleteness of electronic correlation.[83, 84] Villar and coworkers analyzed whether ligand-protein binding enthalpies evaluated by semi-empirical Austin Model 1 (AM1) are sufficient for use in the rational design of new drugs by comparing with B3LYP DFT, and MP2 method.[85] They pointed out that with the exception of cation-π interactions the enthalpies calculated by AM1 correlated well with that by counterpoise-corrected MP2/6-31G(d). However, the structures calculated by AM1 and DFT do not correlated with that calculated by MP2 consistently. Wu and McMahon applied DFT and MP2 to optimize the structures of the most stable isomers of protonated Tyr and ammonia or methylamine and to calculate the enhancement of binding energies due to cation-π interactions. Møller-Plesset perturbation and coupled-cluster methods show that dispersive forces and electrostatic and exchange-repulsion forces play the primary stabilizing role in π-stacked complexes.[84, 86, 87] High-level ab initio calculations, including extrapolation to the MP2 basis set limit and inclusion of a CCSD(T) correction, show that T-shaped and parallel-displaced configurations are virtually isoenergetic in gas phase, with binding energies of -11.46 and $-11.63\,\mathrm{kJ{\cdot}mol^{-1}}$ respectively, whereas the sandwich structure is less stable at $-7.57\,\mathrm{kJ{\cdot}mol^{-1}}$ (5),[84] and substituted benzene dimers bind more strongly than unsubstituted benzene.[88] Hobza and coworkers suggested to model the π-π stacking interactions by MP2 with a medium-sized basis set with a more diffuse polarization function, i.e., MP2/6 31G*(0.25) where exponents of *d* polarization functions are changed into more diffuse 0.25 from 0.8 used in the standard 6-31G* basis.[89–92] Because of its computational efficiency, DFT has been used by several groups to describe π-stacking interactions.[93–97] To attain predicting power similar to high-level ab initio methods, some researchers have combined HF theory and DFT, and using modest basis sets have reproduced the potential energy surface of higher level calculations for a number of instances of π-stacking.[98–100]

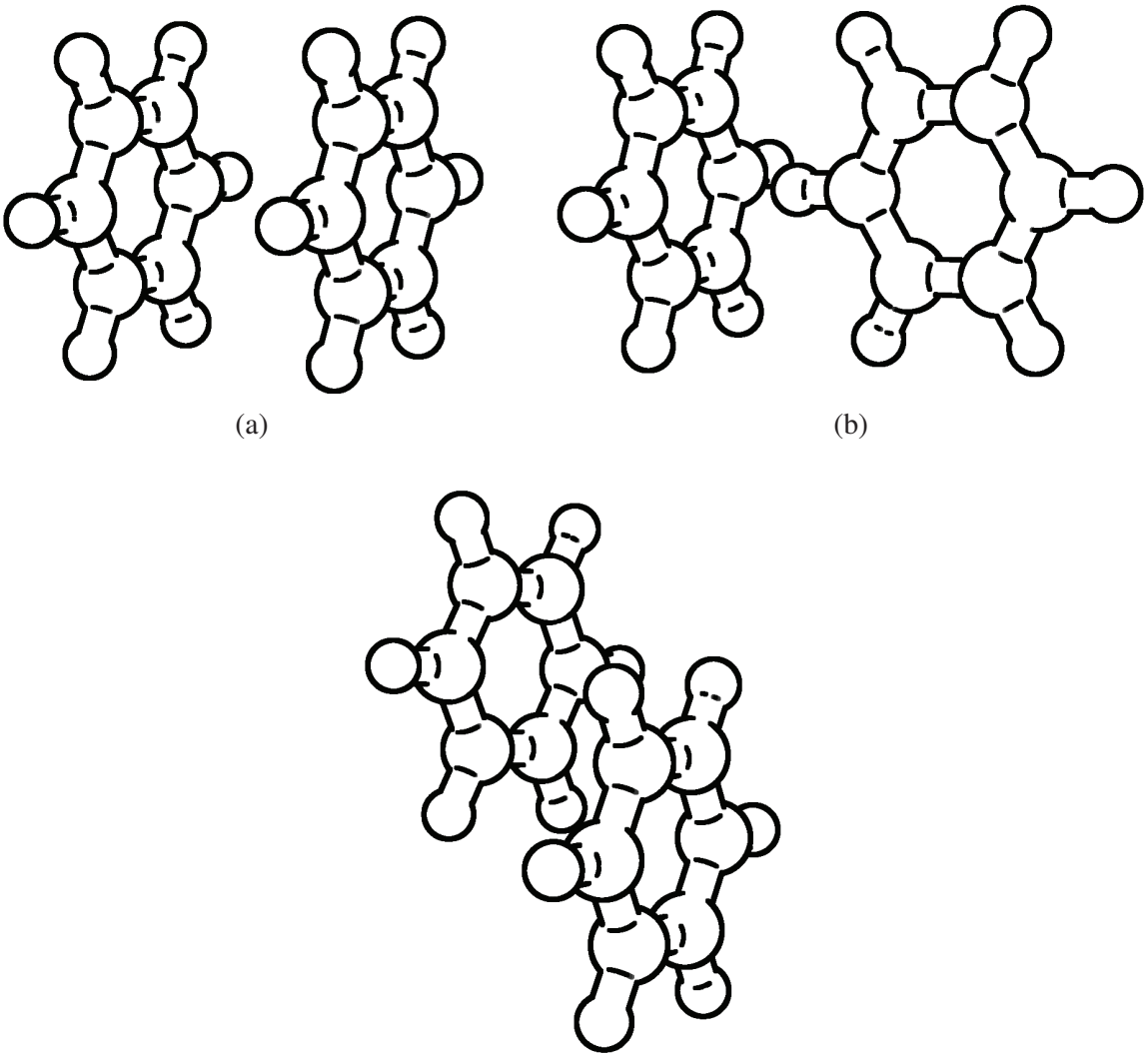


Figure 5: (a) Sandwich, (b) T-shaped, and (c) parallel-displaced configurations of the benzene dimer.[84, 88]

Using QM to calculate molecular properties

It has long been recognized that if one could accurately evaluate the standard free energy change of complexation of biologically active molecules, it would be possible both to gain a deeper understanding of molecular recognition in biology, and to shed light onto the first principles design of pharmaceuticals and other compounds.[101] The currently available computer power does not allow highly accurate QM calculations of free energies, particularly for proteins and ligands in solution. Moreover, usage of QM methods in high-throughput docking is prohibitive. Therefore QM is more suitable to derive models for prediction rather than for the direct evaluation of binding free energies. Classical FFs and QSAR are examples of compromises between accuracy and efficiency.

QM derived FFs

Due to the large chemical space of molecules, FFs do not include all the parameters required for describing drug-like molecules.[102] QM is being used routinely in optimizing geometries, fitting the torsion parameters, and deriving atomic charges for proteins,[103] DNA,[104] and in particular small molecules.[105–108] Spiegel and coworkers developed a new set of FF parameters of platinated moiety via a force matching procedure of the classical forces to ab initio forces obtained from QM/MM trajectories, and extended the classical MD simulation to describe slow converging rearrangement of dinuclear Pt compounds and DNA duplex.[109] Sugiyama and coworkers used DFT calculated partial charges and FF parameters for the atoms near the active site, which are usually significantly polarized, and metal atoms for which FF parameters are not available.[110]

Multipole expansion (ME) is often used in the representation of the molecular electrostatic potential.[30, 111–113] To account for the effects of charge penetration, the point charges, dipoles, quadrupoles, and octupoles in ME model need to be damped. The damping strategies are particular crucial for short-range energies. For example, damping strategies have to be used when ME is applied on calculation of electrostatic potentials or electric fields on van der Waals (vdW) or solvent accessible surface of a molecule.[114] Werneck and coworkers suggested a general methodology

to optimize the damping functions with the ab initio (HF/6-31G** and 6-31G**+) electrostatic potential.[115]

QM-derived partial charges

In the molecular simulations with fixed charge models, the method used to derive partial charges influences the computed physical properties and subsequent docking and scoring significantly.[116, 117] Mobley and coworker compared the hydration free energies of small molecules, whose partial charges are assigned according to different levels of QM, including AM1, HF, DFT, and MP2, by explicit water MD simulations.[118] They found that AM1 bond charge correction method[119] for computing charges works almost as well as any of the more computationally expensive ab initio method. Fischer et al. compared FF-based scoring functions with QM-based scoring functions by computing binding free energies of eleven ligands to the human estrogen receptor subtype α (ERα) and four ligands to the human retinoic acid receptor of isotype γ.[120] They found that the improvement for the complexes with the ERα receptor stemmed from applying classical electrostatic models partial charges derived by fragment molecular orbital (FMO).[121] Illingworth and coworkers implemented QM/MM derived induced charges into a classical framework, redocked 12 difficult protein-ligand complexes with AutoDock,[122] and found that there was no significant improvement in RMSD of the lowest energy structure against the crystal structure but an increment of the largest cluster size.[123] Pasquini and coworkers explained different binding affinities of similar compounds to HIV-1 IN by calculating the partial charges of these compounds and attributed the difference to a poor interaction of the molecules with the divalent metal ions of the active site due to the electron-withdrawing effect.[7] Instead of using fixed and point-charge model, Wang and coworkers calculated solvation free energies of 31 small neutral molecules from QM charge density and continuum dielectric theory (finite-difference PB equation).[124] The QM and PB equations were solved self-consistently until both the charge and reaction field converged. The calculations took into account polarized electronic wave function asymmetric distortion, and spreading out of the electron cloud. In particular, when the solute is treated by QM, part of its electron density penetrates

into the solvent. The experimentally measured solvation free energies of these molecules spanned a range of 25 kcal·mol^{-1}. The authors reported root mean square error of only 1.3 kcal·mol^{-1} upon tuning a single parameter to shift the calculated values.

QM descriptors in QSAR/QSPR models

The information provided by QM is more accurate than FFs, therefore more robust QSAR models and/or QSPR models are expected with QM descriptors.[125] Partial charges are the most common descriptors in QSAR/QSPR models due to their simplicity and informative content. Occhiato and coworkers employed atomic partial charges derived from DFT electrostatic potential in a CoMFA model, with which they designed new 5α-reductase 1 inhibitors.[126] Lepp and Chuman applied LocalSCF calculated atomic charges to build a QSAR model to predict Michaelis-Menten constants, and attained better correlation than classical QSAR descriptors.[127] Wan et al. found that the net charge of the atoms and polarizability correlate with biological activity.[128] Furthermore, they reported that the predictive power of QSAR models derived from DFT charges is higher than from semi-empirical PM3 charges. Besides partial charges, other QM descriptors are commonly used to build QSAR/QSPR models. Yamagami and coworkers used various quantum chemical descriptors, e.g., frontier energy and frontier electron density, which are powerful for describing chemical reactivity.[129] Their CoMFA method shows that the antimutagenic activities are increased by electron-withdrawing substituents and also by hydrogen-bonding between 2-hydroxy group and the receptor. Singh and coworkers developed a QSAR model of derivatives of testosterone with several QM parameters, e.g., absolute hardness and electronegativity.[130] Pasha and coworkers derived QSAR models utilizing various QM descriptors to analyze the factors affecting inhibitory potency for a series of analogues of the MK886 inhibitor of microsomal prostaglandin E2 synthase-1.[131] These QM models indicate that the steric properties, as well as electrostatic and hydrophobic interactions are relevant to the inhibitory potency.

Molecular quantum similarity

Molecular similarity measures have been used in CADD since more than 15 years.[132] Malde and coworkers have investigated boron analogs of natural peptides by QM to find the secondary structural preferences and the impact on stability of different substitutions on boron.[133–135] Recently they have shown that the B(OH)−NH isostere is an interesting surrogate for the peptide bond because of the similar geometry and barrier for rotation around the backbone dihedral angle ω, as well as stability to proteolytic enzymes.[136] Carbó et al. measured the similarity of electron density calculated by QM, and developed a novel QSAR descriptor named molecular quantum similarity measures (MQSM).[137–144] The MQSM relies on the first order electronic density function as molecular descriptor. Before comparing the similarities of electronic density functions, approximated functions[145, 146] and a maximization algorithm are needed to obtain optimal molecular superposition.[147] The MQSM was then used to predict the toxicity[148, 149] and to describe the substituent effect in an aromatic series traditionally described by empirical Hammett equation.[144, 150] MQSM also can be applied for classification of molecules using dendrograms.[151] A further development of MQSM is quantum topological molecular similarity (QTMS),[152] which is based on the definition of distances between molecules in bond critical point space.[153, 154] QTMS is very useful to describe pK_a of molecules as QSAR descriptors.[155, 156] Singh et al. suggested the connection between QTMS and relative bond dissociation enthalpies, and attained good QSAR.[157] Hemmateenejad and Mohajeri used QTMS indices for describing the quantitative effects of molecular electronic environments on the O-methylation kinetic of substituted phenols.[158] Their results revealed that the rate constant of esterification of phenols is highly influenced by the electronic properties of the C_2-C_1-O-H fragment of the parent molecule, which can be considered as frontier bonds in the O-methylation reaction. As shown in these examples, the effects of substitutions are related to the electron density of the bond connecting the scaffolds and the substituents, such that molecular quantum similarity is suitable for studying substitutive effects.

Variational particle number approach for molecular design

Chemical space is the high-dimensional molecular space spanned by the astronomical number of accessible chemical structures. How to sample the chemical space efficiently is always a difficult problem in de novo drug design. In general terms, compound design efforts usually attempt a mapping of a given molecular system to the observable of interest. However in structure-based drug design, the inverse question applies, i.e., which modification of a given compound will result in a desired molecular property. Two independent research groups have recently addressed this question. Lilienfeld and coworkers developed an approach that can explore chemical space in a less heuristic manner by extending the conceptual DFT by the chemical potential for nuclei (alchemical potential).[159] With their approach, they modified a peptidic inhibitor of an anticancer target (human X-chromosome linked inhibitor-of-apoptosis-proteins) into a nonpeptidic inhibitor by optimizing the interaction energy between the inhibitor and the target. Almost at the same time, Wang and coworkers optimized molecular polarizability and hyperpolarizability using a similar method. The main idea of these methods is mapping the discrete chemical structures onto a continuous hypersurface (6). In this case, enumerating the astronomical number of discrete chemical structures can be avoided by a systematic optimization of parameters introduced in the mapping procedure. Up to now, these methods are merely applied for optimizing several molecular properties e.g., polarizability and hyperpolarizability, which can be calculated by QM straightforwardly.[160–166] QSAR/QSPR models may bridge the gap between the properties calculated directly by QM and those that are useful for drug discovery, e.g., high binding affinity and selectivity to the target, good pharmacokinetics and pharmacodynamics, and low toxicity, so that the variational particle number approach might become a routine of structure-based drug design.

Conclusion and outlook

Several QM-based methods already play an important role in many phases of CADD, and will have a stronger impact in the future because of the ever growing computing power and development

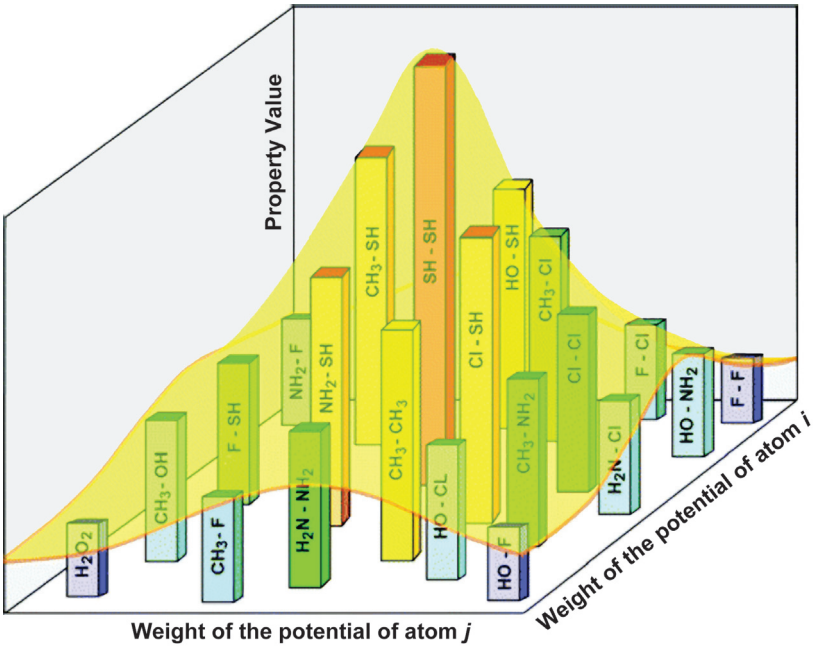


Figure 6: Schematic representation of optimization of molecular properties by a linear combination of atomic potentials. Bar heights represent electronic polarizabilities for candidate structures. The optimization of the property is performed on the smooth (hyper)surface (only two degrees of freedom are denoted). Establishing a well behaved property surface that interpolates among the realizable molecules is a key aspect of the variational particle number approach.[167] This figure is reprinted from Ref.167 with permission of ACS.

of efficient algorithms. The compromise between accuracy and efficiency is a perpetual issue in the applications of QM methods. It is important to select the most appropriate technique at each phase of drug development, and QM methods should be selected only if there is a real advantage with respect to the classical approaches. The initial phase of CADD, e.g., high-throughput docking, which is useful to identify hit compounds,[168] requires full sampling of conformations of the small molecules within the protein binding site. Such extensive sampling calls for approximated energy functions and predicted properties thereof, which are usually calculated by classical FF methods or fast semi-empirical QM methods. In the subsequent phase, hits have to be optimized to leads which does not require extensive sampling but high accuracy because of the small differences in the binding free energy. Therefore QM methods should be applied on the hits to shed light on the energetics of binding. The QM methods are particularly important to capture charge transfer and polarization effects, which are usually pronounced in systems containing metal atoms or charged groups, and/or dispersion forces which play a significant role in the interactions of conjugated π systems. Importantly, before starting CADD it is necessary to evaluate the status of the project, which in turn dictates the number and diversity of molecules to be evaluated and the demand of accuracy, and to select the most appropriate approach accordingly.

Acknowledgement

This work was supported by a grant of the Swiss National Science Foundation to A.C.

References

[1] Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–1818.

[2] Cavalli, A.; Carloni, P.; Recanatini, M. Target-related applications of first principles quantum chemical methods in drug design. *Chem. Rev.* **2006**, *106*, 3497–3519.

[3] Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; WollaCott, A. M.; Westerhoff, L. M.; Merz, K. M.

The role of quantum mechanics in structure-based drug design. *Drug Discov. Today* **2007**, *12*, 725–731.

[4] Peters, M. B.; Raha, K.; Merz, K. M. Quantum mechanics in structure-based drug design. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 370–379.

[5] Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RMI: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27*, 1101–1111.

[6] Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

[7] Pasquini, S.; Mugnaini, C.; Tintori, C.; Botta, M.; Trejos, A.; Arvela, R. K.; Larhed, M.; Witvrouw, M.; Michiels, M.; Christ, F.; Debyser, Z.; Corelli, F. Investigations on the 4-quinolone-3-carboxylic acid motif. 1. Synthesis-activity relationship of a class of human immunodeficiency virus type 1 integrase inhibitors. *J. Med. Chem.* **2008**, *51*, 5125–5129.

[8] Vorobjev, Y. N.; Scheraga, H. A.; Honig, B. Theoretical modeling of electrostatic effects of titratable side-chain groups on protein conformation in a polar ionic solution. 2. pH-induced helix-coil transition of poly-(L-lysine) in water and methanol ionic solutions. *J. Phys. Chem.* **1995**, *99*, 7180–7187.

[9] Vorobjev, Y. N.; Scheraga, H. A.; Hitz, B.; Honig, B. Theoretical modeling of electrostatic effects of titratable side-chain groups on protein conformation in a polar ionic solution. 1. Potential of mean force between charged lysine residues and titration of poly-(L-lysine) in 95% methanol solution. *J. Phys. Chem.* **1994**, *98*, 10940–10948.

[10] Bombasaro, J. A.; Masman, M. F.; Santagata, L. N.; Freile, M. L.; Rodriguez, A. M.; Enriz, R. D. A comprehensive conformational analysis of bullacin B, a potent inhibitor of complex I. Molecular dynamics simulations and ab initio calculations. *J. Phys. Chem. A* **2008**, *112*, 7426–7438.

[11] Butler, K. T.; Luque, F. J.; Barril, X. Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem.* **2009**, *30*, 601–610.

[12] Buback, V.; Mladenovic, M.; Engels, B.; Schirmeister, T. Rational design of improved aziridine-based inhibitors of cysteine proteases. *J. Phys. Chem. B* **2009**, *113*, 5282–5289.

[13] Van der Vaart, A.; Gogonea, V.; Dixon, S. L.; Merz, K. M. Linear scaling molecular orbital calculations of biological systems using the semiempirical divide and conquer method. *J. Comput. Chem.* **2000**, *21*, 1494–1504.

[14] Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. Molecular tailoring approach for simulation of electrostatic properties. *J. Phys. Chem.* **1994**, *98*, 9165–9169.

[15] Dixon, S. L.; Merz, K. M. Semiempirical molecular orbital calculations with linear system size scaling. *J. Chem. Phys.* **1996**, *104*, 6643–6649.

[16] Lee, T. S.; Lewis, J. P.; Yang, W. T. Linear-scaling quantum mechanical calculations of biological molecules: The divide-and-conquer approach. *Comp. Mat. Sci.* **1998**, *12*, 259–277.

[17] Zhang, D. W.; Zhang, J. Z. H. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein-molecule interaction energy. *J. Chem. Phys.* **2003**, *119*, 3599–3605.

[18] Raha, K.; Merz, K. M. Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein–ligand complexes. *J. Med. Chem.* **2005**, *48*, 4558–4575.

[19] Zhou, T.; Huang, D.; Caflisch, A. Is quantum mechanics necessary for predicting binding free energy? *J. Med. Chem.* **2008**, *51*, 4280–4288.

[20] Raha, K.; Merz, K. M. A quantum mechanics-based scoring function: Study of zinc ion-mediated ligand binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.

[21] Huang, D.; Caflisch, A. Efficient evaluation of binding free energy using continuum electrostatics solvation. *J. Med. Chem.* **2004**, *47*, 5791–5797.

[22] Knox, J. E.; Ma, N. L.; Yin, Z.; Patel, S. J.; Wang, W. L.; Chan, W. L.; Rao, K. R. R.; Wang, G.; Ngew, X.; Patel, V.; Beer, D.; Lim, S. P.; Vasudevan, S. G.; Keller, T. H. Peptide inhibitors of West Nile NS3 protease: SAR study of tetrapeptide aldehyde inhibitors. *J. Med. Chem.* **2006**, *49*, 6585–6590.

[23] Yin, Z.; Patel, S. J.; Wang, W. L.; Chan, W. L.; Rao, K. R. R.; Wang, G.; Ngew, X.; Patel, V.; Beer, D.; Knox, J. E.; Ma, N. L.; Ehrhardt, C.; Lim, S. P.; Vasudevan, S. G.; Keller, T. H. Peptide inhibitors of dengue virus NS3 protease. Part 2: SAR study of tetrapeptide aldehyde inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 40–43.

[24] Yin, Z.; Patel, S. J.; Wang, W. L.; Wang, G.; Chan, W. L.; Rao, K. R. R.; Alam, J.; Jeyaraj, D. A.; Ngew, X.; Patel, V.; Beer, D.; Lim, S. P.; Vasudevan, S. G.; Keller, T. H. Peptide inhibitors of dengue virus NS3 protease. Part 1: Warhead. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 36–39.

[25] Dreyer, G. B.; Lambert, D. M.; Meek, T. D.; Carr, T. J.; Tomaszek, T. A.; Fernandez, A. V.; Bartus, H.; Cacciavillani, E.; Hassell, A. M.; Minnich, M.; Petteway, S. R.; Metcalf, B. W.; Lewis, M. Hydroxyethylene isostere inhibitors of human immunodeficiency virus-1 protease - structure activity analysis using enzyme-kinetics, X-ray crystallography, and infected T-cell assays. *Biochemistry (Mosc.)* **1992**, *31*, 6646–6659.

[26] Bramson, H. N. et al. Oxindole-based inhibitors of cyclin-dependent kinase 2 (CDK2): Design, synthesis, enzymatic activities, and X-ray crystallographic analysis. *J. Med. Chem.* **2001**, *44*, 4339–4358.

[27] Gibson, A. E. et al. Probing the ATP ribose-binding domain of cyclin-dependent kinases 1 and 2 with O6 substituted guanine derivatives. *J. Med. Chem.* **2002**, *45*, 3381–3393.

[28] Stewart, J. J. P. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *Int. J. Quantum Chem.* **1996**, *58*, 133–146.

[29] Vasilyev, V.; Bliznyuk, A. Application of semiempirical quantum chemical methods as a scoring function in docking. *Theor. Chem. Acc.* **2004**, *112*, 313–317.

[30] Anikin, N. A.; Anisimov, V. M.; Bugaenko, V. L.; Bobrikov, V. V.; Andreyev, A. M. LocalSCF method for semiempirical quantum-chemical calculation of ultralarge biomolecules. *J. Chem. Phys.* **2004**, *121*, 1266–1270.

[31] Anisimov, V. M.; Bugaenko, V. L.; Bobrikov, V. V. Validation of linear scaling semiempirical LocalSCF method. *J. Chem. Theory Comput.* **2006**, *2*, 1685–1692.

[32] Gao, J. L.; Truhlar, D. G. Quantum mechanical methods for enzyme kinetics. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.

[33] Gao, J. L.; Ma, S. H.; Major, D. T.; Nam, K.; Pu, J. Z.; Truhlar, D. G. Mechanisms and free energies of enzymatic reactions. *Chem. Rev.* **2006**, *106*, 3188–3209.

[34] Cheng, Y. H.; Cheng, X. L.; Radic, Z.; McCammon, J. A. Acetylcholinesterase: Mechanisms of covalent inhibition of H447I mutant determined by computational analyses. *Chem. Biol. Interact.* **2008**, *175*, 196–199.

[35] Hu, H.; Boone, A.; Yang, W. T. Mechanism of OMP decarboxylation in orotidine 5 '-monophosphate decarboxylase. *J. Am. Chem. Soc.* **2008**, *130*, 14493–14503.

[36] Lameira, J.; Alves, C. N.; Moliner, V.; Marti, S.; Kanaan, N.; Tunon, I. A quantum mechanics/molecular mechanics study of the protein–ligand interaction of two potent inhibitors of human O-GlcNAcase: PUGNAc and NAG-thiazoline. *J. Phys. Chem. B* **2008**, *112*, 14260–14266.

[37] Mladenovic, M.; Junold, K.; Fink, R. F.; Thiel, W.; Schirmeister, T.; Engels, B. Atomistic insights into the inhibition of cysteine proteases: First QM/MM calculations clarifying the

regiospecificity and the inhibition potency of epoxide- and aziridine-based inhibitors. *J. Phys. Chem. B* **2008**, *112*, 5458–5469.

[38] Suresh, C. H.; Vargheese, A. M.; Vijayalakshmi, K. P.; Mohan, N.; Koga, N. Role of structural water molecule in HIV protease-inhibitor complexes: A QM/MM study. *J. Comput. Chem.* **2008**, *29*, 1840–1849.

[39] Wu, R. B.; Cao, Z. X. QM/MM study of catalytic methyl transfer by the N-5-Glutamine SAM-dependent methyltransferase and its inhibition by the nitrogen analogue of coenzyme. *J. Comput. Chem.* **2008**, *29*, 350–357.

[40] Kamerlin, S. C. L.; Haranczyk, M.; Warshel, A. Progress in ab initio QM/MM free-energy simulations of electrostatic energies in proteins: accelerated QM/MM studies of pKa, redox reactions and solvation free energies. *J. Phys. Chem. B* **2009**, *113*, 1253–1272.

[41] Lodola, A.; Mor, M.; Sirirak, J.; Mulholland, A. J. Insights into the mechanism and inhibition of fatty acid amide hydrolase from quantum mechanics/molecular mechanics (QM/MM) modelling. *Biochem. Soc. Trans.* **2009**, *37*, 363–367.

[42] Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

[43] Gentilucci, L.; Squassabia, F.; Demarco, R.; Artali, R.; Cardillo, G.; Tolomelli, A.; Spampinato, S.; Bedini, A. Investigation of the interaction between the atypical agonist c[YpwFG] and MOR. *FEBS J.* **2008**, *275*, 2315–2337.

[44] Mukherjee, P.; Desai, P. V.; Srivastava, A.; Tekwani, B. L.; Avery, M. A. Probing the structures of leishmanial farnesyl pyrophosphate synthases: homology modeling and docking studies. *J. Chem. Inf. Model.* **2008**, *48*, 1026–1040.

[45] Fanfrlik, J.; Brynda, J.; Rezac, J.; Hobza, P.; Lepsik, M. Interpretation of protein/ligand

crystal structure using QM/MM calculations: case of HIV-1 protease/metallacarborane complex. *J. Phys. Chem. B* **2008**, *112*, 15094–15102.

[46] Mladenovic, M.; Arnone, M.; Fink, R. F.; Engels, B. Environmental effects on charge densities of biologically active molecules: Do molecule crystal environments indeed approximate protein surroundings? *J. Phys. Chem. B* **2009**, *113*, 5072–5082.

[47] Gleeson, M. P.; Gleeson, D. QM/MM calculations in drug discovery: A useful method for studying binding phenomena? *J. Chem. Inf. Model.* **2009**, *49*, 670–677.

[48] Anisimov, V. M.; Bugaenko, V. L. QM/MM docking method based on the variational finite localized molecular orbital approximation. *J. Comput. Chem.* **2009**, *30*, 784–798.

[49] Gao, X.; Huang, X.; Sun, C. Role of each residue in catalysis in the active site of pyrimidine nucleoside phosphorylase from Bacillus subtilis: A hybrid QM/MM study. *J. Struct. Biol.* **2006**, *154*, 20–26.

[50] Blumberger, J.; Lamoureux, G.; Klein, M. L. Peptide hydrolysis in thermolysin: Ab initio QM/MM investigation of the Glu143-assisted water addition mechanism. *J. Chem. Theory Comput.* **2007**, *3*, 1837–1850.

[51] Zheng, F.; Yang, W.; Ko, M.; Liu, J.; Cho, H.; Gao, D.; Tong, M.; Tai, H.; Woods, J. H.; Zhan, C. Most efficient cocaine hydrolase designed by virtual screening of transition states. *J. Am. Chem. Soc.* **2008**, *130*, 12148–12155.

[52] Wallrapp, F.; Masone, D.; Guallar, V. Electron transfer in the P450cam/PDX complex. The QM/MM e-pathway. *J. Phys. Chem. A* **2008**, *112*, 12989–12994.

[53] Reddy, M. R.; Singh, U. C.; Erion, M. D. Ab initio quantum mechanics-based free energy perturbation method for calculating relative solvation free energies. *J. Comput. Chem.* **2007**, *28*, 491–494.

[54] Magistrato, A.; Ruggerone, P.; Spiegel, K.; Carloni, P.; Reedijk, J. Binding of novel azole-bridged dinuclear Platinum(II) anticancer drugs to DNA: Insights from hybrid QM/MM Molecular Dynamics simulations. *J. Phys. Chem. B* **2006**, *110*, 3604–3613.

[55] Friesner, R. A. Modeling polarization in proteins and protein–ligand complexes: Methods and preliminary results. *Adv. Protein Chem.* **2006**, *72*, 79–104.

[56] Trout, B. L.; Parrinello, M. The dissociation mechanism of H_2O in water studied by first-principles molecular dynamics. *Chem. Phys. Lett.* **1998**, 343–347.

[57] Ridder, L.; Mulholland, A. J. Modeling biotransformation reactions by combined quantum mechanical/molecular mechanical approaches: from structure to activity. *Curr. Top. Med. Chem.* **2003**, *3*, 1241–1256.

[58] Car,; Parrinello, Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

[59] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[60] Ho, M.; Vivo, M. D.; Peraro, M. D.; Klein, M. L. Unraveling the catalytic pathway of metalloenzyme farnesyltransferase through QM/MM computation. *J. Chem. Theory Comput.* **2009**, *5*, 1657–1666.

[61] Feenstra, K. A.; Starikov, E. B.; Urlacher, V. B.; Commandeur, J. N. M.; Vermeulen, N. P. E. Combining substrate dynamics, binding statistics, and energy barriers to rationalize regioselective hydroxylation of octane and lauric acid by CYP102A1 and mutants. *Protein Sci.* **2007**, *16*, 420–431.

[62] Alves, C. N.; Marti, S.; Castillo, R.; Andres, J.; Moliner, V.; Tunon, I.; Silla, E. A quantum mechanic/molecular mechanic study of the wild-type and N155S mutant HIV-1 integrase complexed with diketo acid. *Biophys. J.* **2008**, *94*, 2443–2451.

[63] Reddy, M. R.; Erion, M. D. Relative binding affinities of fructose-1,6-bisphosphatase inhibitors calculated using a quantum mechanics-based free energy perturbation method. *J. Am. Chem. Soc.* **2007**, *129*, 9296–9297.

[64] Sponer, J.; Leszczynski, J.; Hobza, P. Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers* **2001**, *61*, 3–31.

[65] Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S.; Balaz, S. A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands. *J. Med. Chem.* **2005**, *48*, 5437–5447.

[66] Khandelwal, A.; Balaz, S. QM/MM linear response method distinguishes ligand affinities for closely related metalloproteins. *Proteins* **2007**, *69*, 326–339.

[67] Friedman, R.; Caflisch, A. The protonation state of the catalytic aspartates in plasmepsin II. *FEBS Lett.* **2007**, *581*, 4120–4124.

[68] Klein, C. D. P.; Schiffmann, R.; Folkers, G.; Piana, S.; Rothlisberger, U. Protonation states of methionine aminopeptidase and their relevance for inhibitor binding and catalytic activity. *J. Biol. Chem.* **2003**, *278*, 47862–47867.

[69] Fong, P.; McNamara, J. P.; Hillier, I. H.; Bryce, R. A. Assessment of QM/MM scoring functions for molecular docking to HIV-1 protease. *J. Chem. Inf. Model.* **2009**, *49*, 913–924.

[70] Wittayanarakul, K.; Hannongbua, S.; Feig, M. Accurate prediction of protonation state as a prerequisite for reliable MM-PB(GB)SA binding free energy calculations of HIV-1 protease inhibitors. *J. Comput. Chem.* **2008**, *29*, 673–685.

[71] Park, H.; Lee, S. Determination of the active site protonation state of β-secretase from molecular dynamics simulation and docking experiment: Implications for structure-based inhibitor design. *J. Am. Chem. Soc.* **2003**, *125*, 16416–16422.

[72] Rajamani, R.; Reynolds, C. H. Modeling the protonation states of the catalytic aspartates in β-secretase. *J. Med. Chem.* **2004**, *47*, 5159–5166.

[73] Polgar, T.; Keseru, G. M. Virtual screening for β-secretase (BACE1) inhibitors reveals the importance of protonation states at Asp32 and Asp228. *J. Med. Chem.* **2005**, *48*, 3749–3755.

[74] Yu, N.; Hayik, S. A.; Wang, B.; Liao, N.; Reynolds, C. H.; Merz, K. M. Assigning the protonation states of the key aspartates in β-secretase using QM/MM X-ray structure refinement. *J. Chem. Theory Comput.* **2006**, *2*, 1057–1069.

[75] Piana, S.; Sebastiani, D.; Carloni, P.; Parrinello, M. Ab initio molecular dynamics-based assignment of the protonation state of pepstatin A/HIV-1 protease cleavage site. *J. Am. Chem. Soc.* **2001**, *123*, 8730–8737.

[76] Gorfe, A. A.; Caflisch, A. Functional plasticity in the substrate binding site of β-secretase. *Structure* **2005**, *13*, 1487–1498.

[77] Dudev, T.; Lim, C. Metal binding affinity and selectivity in metalloproteins: Insights from computational studies. *Annu. Rev. Biophys.* **2008**, *37*, 97–116.

[78] Seebeck, B.; Reulecke, I.; Kamper, A.; Rarey, M. Modeling of metal interaction geometries for protein–ligand docking. *Proteins* **2008**, *71*, 1237–1254.

[79] Lin, Y. L.; Lim, C. Factors governing the protonation state of Zn-bound histidine in proteins: A DFT/CDM study. *J. Am. Chem. Soc.* **2004**, *126*, 2602–2612.

[80] Dudev, T.; Lim, C. Principles governing Mg, Ca, and Zn binding and selectivity in proteins. *Chem. Rev.* **2003**, *103*, 773–787.

[81] Riccardi, D.; Schaefer, P.; Cui, Q. pKa calculations in solution and proteins with QM/MM free energy perturbation simulations: a quantitative test of QM/MM protocols. *J. Phys. Chem. B* **2005**, *109*, 17715–17733.

[82] Meyer, E. A.; Castellano, R. K.; Diederich, F. Interactions with aromatic rings in chemical and biological recognition. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210–1250.

[83] Hobza, P.; Selzle, H. L.; Schlag, E. W. Potential energy surface for the benzene dimer. Results of ab initio CCSD(T) calculations show two nearly isoenergetic structures: T-shaped and parallel-displaced. *J. Phys. Chem.* **1996**, *100*, 18790–18794.

[84] Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. Estimates of the ab initio limit for π–π interactions: The benzene dimer. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10893.

[85] Villar, R.; Gil, M. J.; Garcia, J. I.; Martinez-Merino, V. Are AM1 ligand-protein binding enthalpies good enough for use in the rational design of new drugs? *J. Comput. Chem.* **2005**, *26*, 1347–1358.

[86] Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. Molecular clusters of π-systems: Theoretical studies of structures, spectra, and origin of interaction energies. *Chem. Rev.* **2000**, *100*, 4145–4185.

[87] Wu, R. H.; McMahon, T. B. Investigation of cation–π interactions in biological systems. *J. Am. Chem. Soc.* **2008**, *130*, 12554–12555.

[88] Sinnokrot, M. O.; Sherrill, C. D. Substituent effects in π–π interactions: Sandwich and T-shaped configurations. *J. Am. Chem. Soc.* **2004**, *126*, 7690–7697.

[89] Sponer, J.; Gabb, H. A.; Leszczynski, J.; Hobza, P. Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: A quantum-chemical study. *Biophys. J.* **1997**, *73*, 76–87.

[90] Hobza, P.; Sponer, J. Significant structural deformation of nucleic acid bases in stacked base pairs: an ab initio study beyond Hartree-Fock. *Chem. Phys. Lett.* **1998**, *288*, 7–14.

[91] Reha, D.; Kabelac, M.; Ryjacek, F.; Sponer, J.; Sponer, J. E.; Elstner, M.; Suhai, S.; Hobza, P. Intercalators. 1. Nature of stacking interactions between intercalators (ethidium, daunomycin, ellipticine, and 4 ',6-diaminide-2-phenylindole) and DNA base pairs. Ab initio quantum chemical, density functional theory, and empirical potential study. *J. Am. Chem. Soc.* **2002**, *124*, 3366–3376.

[92] Jurecka, P.; Hobza, P. True stabilization energies for the optimal planar hydrogen-bonded and stacked structures of guanine center dot center dot center dot cytosine, adenine center dot center dot center dot thymine, and their 9-and 1-methyl derivatives: Complete basis set calculations at the MP2 and CCSD(T) levels and comparison with experiment. *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613.

[93] Meijer, E. J.; Sprik, M. A density-functional study of the intermolecular interactions of benzene. *J. Chem. Phys.* **1996**, *105*, 8684–8689.

[94] Ye, X. Y.; Li, Z. H.; Wang, W. N.; Fan, K. N.; Xu, W.; Hua, Z. Y. The parallel π–π stacking: a model study with MP2 and DFT methods. *Chem. Phys. Lett.* **2004**, *397*, 56–61.

[95] Xu, X.; Goddard, W. A. The X3LYP extended density functional for accurate descriptions of nonbond interactions, spin states, and thermochemical properties. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 2673–2677.

[96] Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. Application of 25 density functionals to dispersion-bound homomolecular dimers. *Chem. Phys. Lett.* **2004**, *394*, 334–338.

[97] Cerny, J.; Hobza, P. The X3LYP extended density functional accurately describes H-bonding but fails completely for stacking. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624–1626.

[98] Perez-Jorda, J. M.; San-Fabian, E.; Perez-Jimenez, A. J. Density-functional study of van der Waals forces on rare-gas diatomics: Hartree-Fock exchange. *J. Chem. Phys.* **1999**, *110*, 1916–1920.

[99] Walsh, T. R. Exact exchange and Wilson-Levy correlation: a pragmatic device for studying complex weakly-bonded systems. *Phys. Chem. Chem. Phys.* **2005**, *7*, 443–451.

[100] Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. Hybrid density functional theory for pi-stacking interactions: Application to benzenes, pyridines, and DNA bases. *J. Comput. Chem.* **2006**, *27*, 491–504.

[101] McCammon, J. A. Free energy calculations in rational drug design. *Abstracts of Papers of the American Chemical Society* **2004**, *227*, U896–U896.

[102] Curioni, A.; Mordasini, T.; Andreoni, W. Enhancing the accuracy of virtual screening: molecular dynamics with quantum-refined force fields. *J. Comput. Aided Mol. Des.* **2004**, *18*, 773–784.

[103] Ponder, J. W.; Case, D. A. Force fields for protein simulations. *Protein Simulations* **2003**, *66*, 27–85.

[104] Ryjacek, F.; Kubar, T.; Hobza, P. New parameterization of the Cornell et al. empirical force field covering amino group nonplanarity in nucleic acid bases. *J. Comput. Chem.* **2003**, *24*, 1891–1901.

[105] Maple, J. R.; Dinur, U.; Hagler, A. T. Derivation of force fields for molecular mechanics and dynamics from ab initio energy surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 5350–5354.

[106] Palmo, K.; Mannfors, B.; Mirkin, N. G.; Krimm, S. Potential energy functions: From consistent force fields to spectroscopically determined polarizable force fields. *Biopolymers* **2003**, *68*, 383–394.

[107] Maurer, P.; Laio, A.; Hugosson, H. W.; Colombo, M. C.; Rothlisberger, U. Automated parametrization of biomolecular force fields from quantum mechanics/molecular mechanics (QM/MM) simulations through force matching. *J. Chem. Theory Comput.* **2007**, *3*, 628–639.

[108] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2009**, in press.

[109] Spiegel, K.; Magistrato, A.; Maurer, P.; Ruggerone, P.; Rothlisberger, U.; Carloni, P.; Reedijk, J.; Klein, M. L. Parameterization of azole-bridged dinuclear platinum anticancer drugs via a QM/MM force matching procedure. *J. Comput. Chem.* **2007**, *29*, 38–49.

[110] Sugiyama, A.; Takamatsu, Y.; Nishikawa, K.; Nagao, H.; Nishikawa, K. Docking stability and electronic structure of azurin-cytochrome c551 complex system. *Int. J. Quantum Chem.* **2006**, *106*, 3071–3078.

[111] Dykstra, C. E. Electrostatic interaction potentials in molecular-force fields. *Chem. Rev.* **1993**, *93*, 2339–2353.

[112] Narayszabo, G.; Ferenczy, G. G. Molecular electrostatics. *Chem. Rev.* **1995**, *95*, 829–847.

[113] Engkvist, O.; Astrand, P. O.; Karlstrom, G. Accurate intermolecular potentials obtained from molecular wave functions: Bridging the gap between quantum chemistry and molecular simulations. *Chem. Rev.* **2000**, *100*, 4087–4108.

[114] Dardenne, L. E.; Werneck, A. S.; Neto, M. D.; Bisch, P. M. Electrostatic properties in the catalytic site of papain: A possible regulatory mechanism for the reactivity of the ion pair. *Proteins* **2003**, *52*, 236–253.

[115] Werneck, A. S.; Filho, T. M. R.; Dardenne, L. E. General methodology to optimize damping functions to account for charge penetration effects in electrostatic calculations using multicentered multipolar expansions. *J. Phys. Chem. A* **2008**, *112*, 268–280.

[116] Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. Importance of accurate charges in molecular

docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem.* **2005**, *26*, 915–931.

[117] Kumar, S.; Jaller, D.; Patel, B.; LaLonde, J. M.; DuHadaway, J. B.; Malachowski, W. P.; Prendergast, G. C.; Muller, A. J. Structure based development of phenylimidazole-derived inhibitors of indoleamine 2,3-dioxygenase. *J. Med. Chem.* **2008**, *51*, 4968–4977.

[118] Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.

[119] Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

[120] Fischer, B.; Fukuzawa, K.; Wenzel, W. Receptor-specific scoring functions derived from quantum chemical models improve affinity estimates for in-silico drug discovery. *Proteins* **2008**, *70*, 1264–1273.

[121] Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* **1999**, *313*, 701–706.

[122] Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195–202.

[123] Illingworth, C. J. R.; Morris, G. M.; Parkes, K. E. B.; Snell, C. R.; Reynolds, C. A. Assessing the role of polarization in docking. *J. Phys. Chem. A* **2008**, *112*, 12157–12163.

[124] Wang, M. L.; Wong, C. F. Calculation of solvation free energy from quantum mechanical charge density and continuum dielectric theory. *J. Phys. Chem. A* **2006**, *110*, 4873–4879.

[125] Gacche, R.; Khsirsagar, M.; Kamble, S.; Bandgar, B.; Dhole, N.; Shisode, K.; Chaudhari, A. Antioxidant and anti-inflammatory related activities of selected synthetic chalcones: structure-

activity relationship studies using computational tools. *Chem. Pharm. Bull. (Tokyo)* **2008**, *56*, 897–901.

[126] Occhiato, E. G.; Ferrali, A.; Menchi, G.; Guarna, A.; Danza, G.; Comerci, A.; Mancina, R.; Serio, M.; Garotta, G.; Cavalli, A.; De Vivo, M.; Recanatini, M. Synthesis, biological activity, and three-dimensional quantitative structure-activity relationship model for a series of benzo[c]quinolizin-3-ones, nonsteroidal inhibitors of human steroid 5 α-reductase 1. *J. Med. Chem.* **2004**, *47*, 3546–3560.

[127] Lepp, Z.; Chuman, H. Connecting traditional QSAR and molecular simulations of papain hydrolysis - importance of charge transfer. *Bioorg. Med. Chem.* **2005**, *13*, 3093–3105.

[128] Wan, J.; Zhang, L.; Yang, G. F.; Zhan, C. G. Quantitative structure-activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: A study of quantum chemical descriptors from density functional theory. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2099–2105.

[129] Yamagami, C.; Motohashi, N.; Akamatsu, M. Quantum chemical- and 3-D-QSAR (CoMFA) studies of benzalacetones and 1,1,1-trifluoro-4-phenyl-3-buten-2-ones. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 2281–2285.

[130] Singh, P. P.; Srivastava, H. K.; Pasha, F. A. DFT-based QSAR study of testosterone and its derivatives. *Bioorg. Med. Chem.* **2004**, *12*, 171–177.

[131] Pasha, F. A.; Muddassar, M.; Jung, H.; Yang, B.; Lee, C.; Oh, J. S.; Cho, S. J.; Cho, H. QM and pharmacophore based 3D-QSAR of MK886 analogues against mPGES-1. *Bull. Korean Chem. Soc.* **2008**, *29*, 647–655.

[132] Good, A. C.; So, S. S.; Richards, W. G. Structure-activity-relationships from molecular similarity-matrices. *J. Med. Chem.* **1993**, *36*, 433–438.

[133] Malde, A.; Khedkar, S.; Coutinho, E.; Saran, A. Geometry, transition states, and vibrational spectra of boron isostere of N-methylacetamide by ab initio calculations. *Int. J. Quantum Chem.* **2005**, *102*, 734–742.

[134] Malde, A. K.; Khedkar, S. A.; Coutinho, E. C. Stationary points on the PES of N-methoxy peptides and their boron isosteres: An ab initio study. *J. Chem. Theory Comput.* **2006**, *2*, 1664–1674.

[135] Malde, A. K.; Khedkar, S. A.; Coutinho, E. C. Isosteres of peptides: boron analogs as dipolar forms of alpha-amino acids - a theoretical study. *J. Phys. Org. Chem.* **2007**, *20*, 151–160.

[136] Malde, A. K.; Khedkar, S. A.; Coutinho, E. C. The B(OH)–NH analog is a surrogate for the amide bond (CO–NH) in peptides: An ab initio study. *J. Chem. Theory Comput.* **2007**, *3*, 619–627.

[137] Carbó, R.; Besalu, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR). *J. Math. Chem.* **1995**, *18*, 237–246.

[138] Fradera, X.; Amat, L.; Besalu, E.; Carbó-Dorca, R. Application of molecular quantum similarity to QSAR. *Quantitative Structure-Activity Relationships* **1997**, *16*, 25–32.

[139] Lobato, M.; Amat, L.; Besalu, E.; Carbó-Dorca, R. Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quantitative Structure-Activity Relationships* **1997**, *16*, 465–472.

[140] Amat, L.; Robert, D.; Besalu, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: An antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624–631.

[141] Besalu, E.; Girones, X.; Amat, L.; Carbó-Dorca, R. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res.* **2002**, *35*, 289–295.

[142] Bultinck, P.; Carbó-Dorca, R.; Van Alsenoy, C. Quality of approximate electron densities and internal consistency of molecular alignment algorithms in molecular quantum similarity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1208–1217.

[143] Bultinck, P.; Kuppens, T.; Girone, X.; Carbó-Dorca, R. Quantum similarity superposition algorithm (QSSA): A consistent scheme for molecular alignment and molecular similarity based on quantum chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1143–1150.

[144] Girones, X.; Carbó-Dorca, R.; Ponec, R. Molecular basis of LFER. Modeling of the electronic substituent effect using fragment quantum self-similarity measures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2033–2038.

[145] Constans, P.; Carbó, R. Atomic shell approximation - electron-density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046–1053.

[146] Amat, L.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: First order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023–2039.

[147] Constans, P.; Amat, L.; Carbó-Dorca, R. Toward a global maximization of the molecular similarity function: Superposition of two molecules. *J. Comput. Chem.* **1997**, *18*, 826–846.

[148] Gallegos, A.; Robert, D.; Girones, X.; Carbó-Dorca, R. Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. *J. Comput. Aided Mol. Des.* **2001**, *15*, 67–80.

[149] Girones, X.; Carbó-Dorca, R. Modelling toxicity using molecular quantum similarity measures. *QSAR Comb. Sci.* **2006**, *25*, 579–589.

[150] Girones, X.; Ponec, R. Molecular quantum similarity measures from Fermi hole densities: Modeling Hammett sigma constants. *J. Chem. Inf. Model.* **2006**, *46*, 1388–1393.

[151] Bultinck, P.; Carbó-Dorca, R. Molecular quantum similarity matrix based clustering of molecules using dendrograms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 170–177.

[152] O'Brien, S. E.; Popelier, P. L. A. Quantum molecular similarity. 3. QTMS descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 764–775.

[153] Popelier, P. L. A. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem. A* **1999**, *103*, 2883–2890.

[154] O'Brien, S. E.; Popelier, P. L. A. Quantum molecular similarity. Part 2: The relation between properties in BCP space and bond length. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **1999**, *77*, 28–36.

[155] Chaudry, U. A.; Popelier, P. L. A. Estimation of pK(a) using quantum topological molecular similarity descriptors: Application to carboxylic acids, anilines and phenols. *J. Org. Chem.* **2004**, *69*, 233–241.

[156] Esteki, M.; Hemmateenejad, B.; Khayamian, T.; Mohajeri, A. Multi-way analysis of quantum topological molecular similarity descriptors for modeling acidity constant of some phenolic compounds. *Chem. Biol. Drug Des.* **2007**, *70*, 413–423.

[157] Singh, N.; Loader, R. J.; O'Malley, P. J.; Popelier, P. L. A. Computation of relative bond dissociation enthalpies (Delta BDE) of phenolic antioxidants from quantum topological molecular similarity (QTMS). *J. Phys. Chem. A* **2006**, *110*, 6498–6503.

[158] Hemmateenejad, B.; Mohajeri, A. Application of quantum topological molecular similarity descriptors in QSPR study of the O-methylation of substituted phenols. *J. Comput. Chem.* **2008**, *29*, 266–274.

[159] von Lilienfeld, O. A.; Lins, R. D.; Rothlisberger, U. Variational particle number approach for rational compound design. *Phys. Rev. Lett.* **2005**, *95*, Doi 10.1103.

[160] von Lilienfeld, O. A.; Tuckerman, M. E. Molecular grand-canonical ensemble density functional theory and exploration of chemical space. *J. Chem. Phys.* **2006**, *125*, 154104–154113.

[161] Keinan, S.; Hu, X. Q.; Beratan, D. N.; Yang, W. T. Designing molecules with optimal properties using the linear combination of atomic potentials approach in an AM1 semiempirical framework. *J. Phys. Chem. A* **2007**, *111*, 176–181.

[162] von Lilienfeld, O. A.; Tuckerman, M. E. Alchemical variations of intermolecular energies according to molecular grand-canonical ensemble density functional theory. *J. Chem. Theory Comput.* **2007**, *3*, 1083–1090.

[163] Balamurugan, D.; Yang, W. T.; Beratan, D. N. Exploring chemical space with discrete, gradient, and hybrid optimization methods. *J. Chem. Phys.* **2008**, *129*, 174105–174113.

[164] Hu, X. Q.; Beratan, D. N.; Yang, W. T. A gradient-directed Monte Carlo approach to molecular design. *J. Chem. Phys.* **2008**, *129*, 64102–64110.

[165] Keinan, S.; Paquette, W. D.; Skoko, J. J.; Beratan, D. N.; Yang, W. T.; Shinde, S.; Johnston, P. A.; Lazo, J. S.; Wipf, P. Computational design, synthesis and biological evaluation of para-quinone-based inhibitors for redox regulation of the dual-specificity phosphatase Cdc25B. *Org. Biomol. Chem.* **2008**, *6*, 3256–3263.

[166] Keinan, S.; Therien, M. J.; Beratan, D. N.; Yang, W. T. Molecular design of porphyrin-based nonlinear optical materials. *J. Phys. Chem. A* **2008**, *112*, 12203–12207.

[167] Wang, M. L.; Hu, X. Q.; Beratan, D. N.; Yang, W. T. Designing molecules by optimizing potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228–3232.

[168] Huang, D.; Caflisch, A. Library screening by fragment-based docking. *J. Molec. Recogn.* **2009**, in press.

Chapter 2

Is Quantum Mechanics Necessary for Predicting Binding Free Energy?

Zhou, T.; Huang, D.; Caflisch A. *J. Med. Chem.* **2008**, *51 (14)*, 4280–4288

Is Quantum Mechanics Necessary for Predicting Binding Free Energy?

Ting Zhou, Danzhi Huang,* and Amedeo Caflisch*

Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received March 6, 2008

To take into account polarization effects, the linear interaction energy model with continuum electrostatic solvation (LIECE) is supplemented by the linear-scaling semiempirical quantum mechanical calculation of the intermolecular electrostatic energy (QMLIECE). QMLIECE and LIECE are compared on three enzymes belonging to different classes: the West Nile virus NS3 serine protease (WNV PR), the aspartic protease of the human immunodeficiency virus (HIV-1 PR), and the human cyclin-dependent kinase 2 (CDK2). QMLIECE is superior for 44 peptidic inhibitors of WNV PR because of the different amount of polarization due to the broad range of formal charges of the inhibitors (from 0 to 3). On the other hand, QMLIECE and LIECE show similar accuracy for 24 peptidic inhibitors of HIV-1 PR (20 neutral and 4 with one formal charge) and for 73 CDK2 inhibitors (all neutral). These results indicate that quantum mechanics is essential when the inhibitor/protein complexes have highly variable charge−charge interactions.

1. Introduction

Accurate methods for computing the binding affinity between small molecules and proteins are needed for drug discovery and design.[1] Approaches based on ab initio quantum mechanics (QM) are rigorous but slow for the studies of macromolecules of biological interest. In order to accelerate QM calculations the hybrid QM/molecular mechanics method[2,3] has been developed for the study of enzyme catalysis.[4–7] In addition linear-scaling QM approaches[8–11] have also been applied extensively for the evaluation of binding affinity between small molecules and proteins.[12,13] Because of their first principle nature, both the time-consuming ab initio methods[14] and fast semiempirical methods[15,16] do not suffer of the approximation inherent to the ball-spring description and the fix-charge approach used in the force field method. Raha and Merz[17] developed a semiempirical/linearly scaling QM-based scoring function and studied the ion-mediated ligand binding processes. They pointed out that QM is needed for metal-containing system because the ill-defined atom types of metal atoms in most of the force field parameters cannot describe the nature of the interactions between a small molecule and a metal ion in the active site. Nevertheless, even with the fast semiempirical/linearly scaling methods, QM approaches are still time-consuming compared with force field based methods especially for high throughput docking.[18,19] Moreover, most of the QM methods significantly underestimate the weak London dispersion forces which require highly correlated theoretical levels and large basis sets.[20] These weak forces, however, play a major role in the hydrophobic effect, molecular recognition, and ligand binding.[21–23] Therefore, it is important to find an optimal compromise between accuracy and efficiency in binding affinity calculation for high-throughput docking campaigns of multimillion library of compounds.

Recently, the linear interaction energy model with continuum electrostatic solvation (LIECE[a])[24] has been successfully applied in high-throughput docking resulting in the discovery of inhibitors of proteases[25,26] and kinases.[27] LIECE is about 2 orders of magnitude faster than the original LIE[28–30] method because molecular dynamics (MD) sampling is replaced by a simple energy minimization. In this paper, LIECE is further improved by using a linearly scaling semiempirical QM method[9] to calculate the electrostatic interaction energy between the ligand and the protein, and the new approach is termed

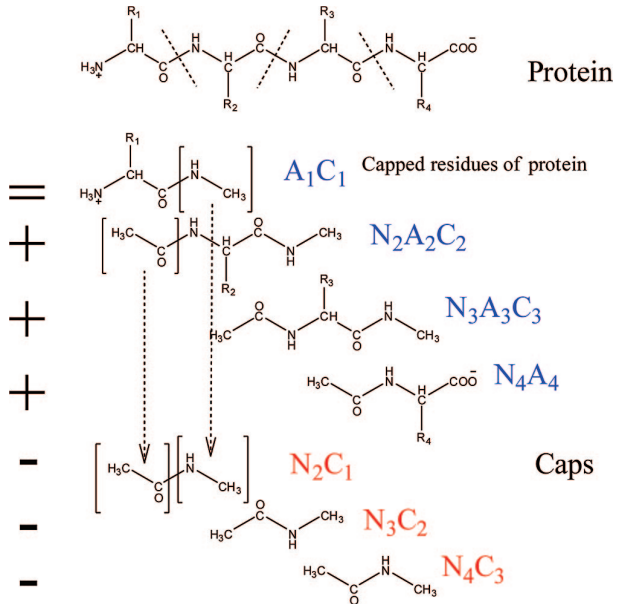


Figure 1. Divide and conquer protocol[9] for calculation of quantum mechanical interaction energy between a protein and a small molecule (ligand). The interaction energy between a protein with *m* residues and the ligand is decomposed into

$$E_{\text{ligand-protein}} = E_{\text{ligand-A}_1\text{C}_1} + E_{\text{ligand-N}_2\text{A}_2\text{C}_2} + \ldots + E_{\text{ligand-N}_{m-1}\text{A}_{m-1}\text{C}_{m-1}} + E_{\text{ligand-N}_m\text{A}_m} - E_{\text{ligand-N}_2\text{C}_1} - E_{\text{ligand-N}_3\text{C}_2} - \ldots - E_{\text{ligand-N}_m\text{C}_{m-1}}$$

where N_i and C_i are N-terminal and C-terminal cap, respectively, of residue A_i. The fragments with blue names are protein residues with conjugate caps,[10] while the ones with red names are pure "caps" that have to be subtracted to remove the duplication in energy calculation.

* To whom correspondence should be addressed. (D.H.) Phone: (+41 44) 635 55 21. Fax: (+41 44) 635 68 62. E-mail: huang@bioc.uzh.ch. (A.C.) Phone: (+41 44) 635 55 21. Fax: (+41 44) 635 68 62. E-mail: caflisch@bioc.uzh.ch.

[a] Abbreviations: LIECE, linear interaction energy model with continuum electrostatic solvation; QMLIECE, quantum mechanical linear interaction energy model with continuum electrostatic solvation; CDK2, cyclin-dependent kinase 2; HIV-1 PR, human immunodeficiency virus protease; WNV PR, West Nile virus NS3 serine protease.

10.1021/jm800242q CCC: $40.75
Published on Web 06/25/2008

Table 1. 44 Peptidic Inhibitors of WNV PR Tested at the Novartis Institute for Tropical Diseases[34–36,a]

ID	structure	no. of formal charges	IC_{50} (μm)	ΔG (kcal/mol)
1	Bz-Nle-Lys-Arg-Arg-H	3	4.1	−7.39
2	Bz-Nle-Lys-Lys(Z)-Arg-H	2	99.5	−5.49
3	Bz-Nle-Lys-Gln-Arg-H	2	1.7	−7.90
4	Bz-Nle-Lys-Lys-Arg-H	3	1.9	−7.86
5	Ac-Lys-Lys-Arg-H	3	**0.4**	−8.84
6	Bz-Lys-Arg-Arg-H	3	1.5	−7.99
7	Bz-Lys-Lys(Tos)-Arg-H	2	117.9	−5.39
8	Ac-Lys-Lys(Tos)-Arg-H	2	**463.4**	−4.58
9	Ac-Lys-Lys(Bz)-Arg-H	2	116.5	−5.40
10	Bz-Lys-(p-Me)Phe-Arg-H	2	194.3	−5.09
11	Bz-Lys-Lys(Bz)-Arg-H	2	68.1	−5.72
12	indole-Lys-Arg-Arg-H	3	2.4	−7.72
13	Bz-Lys-Asn-Arg-H	2	71.8	−5.69
14	Bz-Nle-Ala-Arg-Arg-H	2	3.8	−7.44
15	Bz-Ala-Lys-Arg-Arg-H	3	0.7	−8.45
16	Bz-Nle-Lys-Arg-Phe-H	2	109.8	−5.43
17	Bz-Nle-Lys-Phe-Arg-H	2	108.0	−5.44
18	Bz-Nle-Phe-Arg-Arg-H	2	4.2	−7.38
19	Bz-Phe-Lys-Arg-Arg-H	3	1.2	−8.14
20	Bz-Lys-Arg-Tyr-H	2	14.6	−6.64
21	Ac-KRR-H	3	0.5	−8.60
22	pyridine-KRR-H	3	0.8	−8.40
23	isoquinoline-KRR-H	3	0.6	−8.56
24	pyrazine-Lys-Arg-Arg-H	3	1.1	−8.18
25	3-pyridyl-KRR-H	3	1.0	−8.24
26	Bzl-Nle-Lys-Arg-(4-CN)-Phe-H	2	62.0	−5.77
27	Bzl-Nle-Lys-Arg-Trp-H	2	10.0	−6.86
28	Bz-Nle-Lys-Arg-Lys-H	3	57.7	−5.82
29	BZ-Nle-Lys-Arg-(4-guanidinyl)-Phe-H	3	11.8	−6.76
30	Bz-Nle-Lys-Arg-His-H	2	43.1	−5.99
31	Bz-Nle-Lys-Arg-Phg-H	2	90.9	−5.55
32	Bz-Arg-Arg-H	2	3.9	−7.42
33	Bz-Lys(Z)-Arg-H	1	436.3	−4.61
34	Bz-Lys-Arg-H	2	1.4	−8.03
35	Bz-Arg-Lys-H	2	57.5	−5.82
36	Bz-Lys(Z)-(2-napthyl)Ala-H	0	15.9	−6.59
37	Bz-Lys(Z)-Tyr(Bn)-H	0	17.2	−6.54
38	Bz-Tyr(Bn)-(p-Me)Ph-H	0	12.7	−6.72
39	Bz-Lys(Z)-(p-NH-1-isoquinoline)Phe-H	0	18.0	−6.51
40	Bz-Lys(Z)-(p-NH-1-indole)Phe-H	0	38.2	−6.06
41	Bz-Lys-Lys(Z)-(p-Me)Phe-H	1	43.3	−5.99
42	Bz-Lys-Arg-Phe-H	2	71.1	−5.69
43	Bz-Lys-Arg-(P-Me)Phe-H	2	17.7	−6.52
44	Bz-Lys-Arg-Tyr(Bn)-H	2	11.8	−6.76

[a] Weakest and strongest affinities are in bold.

QMLIECE. The LIECE and QMLIECE models are tested on three enzyme/inhibitor systems: the West Nile virus NS3 protease (WNV PR, a serine protease), the HIV-1 protease (HIV-1 PR, an aspartic protease), and the human cyclin-dependent kinase 2 (CDK2). Because of the large variability of charge−charge interactions in the complexes of WNV PR with the 44 peptidic inhibitors (having between 0 and 3 positively charged side chains), the use of QM is necessary to capture polarization effects,[31,32] which are neglected in fixed-charge approximation of force field based methods. On the other hand, QM and force field methods show similar accuracy for the binding affinity evaluation of mainly neutral inhibitors of HIV-1 PR and CDK2.

2. Method

Preparation of WNV NS3-NS2B Protease. The coordinates of WNV PR in complex with the substrate-based inhibitor benzoyl-norleucine-lysine-arginine-arginine-aldehyde (Bz-Nle-Lys-Arg-Arg-H) were downloaded from the PDB database (PDB entry 2FP7).[33] All water molecules were removed. The spurious termini at the segments missing in the X-ray structure (residues 28−32 in chain B) were neutralized by a $-COCH_3$ and a $-NHCH_3$ group at the N-terminus and C-terminus, respectively. The 44 peptidic inhibitors of WNV PR used in this study include Bz-Nle-Lys-Arg-Arg-H (IC_{50} = 4.1 μM) and a series of 43 related inhibitors with an aldehyde warhead (IC_{50} values ranging from 0.4 to 463 μM) synthesized in the same laboratory and tested all with the same enzymatic assay (Table 1).[34–36] The initial binding conformations were modeled manually according to the binding mode of Bz-Nle-Lys-Arg-Arg-H because all inhibitors have similar backbone structure and are covalently bound to the Ser135 side chain by an ester linkage.

For interaction energy calculation, the ester bond between protein and inhibitor and the adjacent −OH group of Ser135 and −C(H)OH group of inhibitor were removed to avoid the artificial crash. The resulting empty valencies on both protein and inhibitors were filled with hydrogen atoms.

Preparation of HIV-1 PR and CDK2. The coordinates of the 24 complexes of HIV-1 PR (PDB code 1AAQ) with the inhibitors tested by Dreyer and co-workers[37] were available from a previous study.[24] The coordinates of the 73 complexes of

Table 2. Energy Components for WNV PR Peptidic Inhibitors[34–36,a]

ID	ΔE_{vdW}	$\Delta E_{elec,coul}$	ΔE_{QM}	$\Delta G^{solvation}$	no. of formal charges
1	−45.2	−898.0	−856.8	846.9	3
2	−53.2	−603.1	−511.1	608.9	2
3	−41.1	−601.1	−516.6	566.1	2
4	−40.9	−880.5	−874.2	840.1	3
5	−38.4	−926.0	−950.3	883.2	3
6	−41.2	−899.4	−860.7	842.0	3
7	−46.3	−621.8	−530.3	616.5	2
8	−45.8	−617.6	−530.2	616.0	2
9	−48.3	−593.8	−513.1	601.3	2
10	−43.7	−583.8	−511.5	558.0	2
11	−47.2	−596.7	−503.7	595.3	2
12	−36.9	−881.8	−874.0	839.0	3
13	−35.7	−589.4	−506.0	556.0	2
14	−46.3	−620.5	−530.6	565.0	2
15	−44.9	−898.5	−862.2	845.9	3
16	−54.2	−654.5	−565.1	621.2	2
17	−43.5	−580.6	−510.4	556.6	2
18	−53.9	−623.6	−531.1	568.3	2
19	−45.9	−898.1	−860.2	846.0	3
20	−52.8	−657.1	−565.0	616.9	2
21	−39.3	−898.7	−869.7	847.2	3
22	−40.9	−905.7	−870.4	847.6	3
23	−41.0	−904.3	−868.3	847.8	3
24	−40.4	−900.4	−862.9	842.5	3
25	−40.9	−901.8	−866.5	845.4	3
26	−52.3	−661.8	−573.2	623.1	2
27	−56.4	−650.8	−559.7	619.0	2
28	−45.8	−950.8	−870.4	906.6	3
29	−47.7	−903.9	−851.7	867.8	3
30	−50.3	−656.0	−554.7	619.3	2
31	−52.6	−660.0	−556.5	629.5	2
32	−41.7	−628.0	−539.2	565.0	2
33	−49.7	−362.3	−269.2	357.1	1
34	−36.1	−649.8	−566.5	595.6	2
35	−37.3	−672.9	−566.3	615.7	2
36	−52.9	−111.5	−34.2	116.3	0
37	−59.6	−113.0	−35.4	121.3	0
38	−54.0	−70.9	−11.6	81.1	0
39	−58.0	−124.5	−44.2	136.4	0
40	−58.8	−121.4	−43.4	124.4	0
41	−60.1	−350.5	−260.1	347.3	1
42	−48.4	−654.1	−553.8	621.3	2
43	−47.1	−651.3	−552.4	616.2	2
44	−49.5	−653.5	−547.7	616.3	2

[a] All energy values are in kcal·mol^{-1}.

Table 3. Energy Components of HIV-1 PR Peptidic Inhibitors[37,a]

ID	ΔE_{vdW}	$\Delta E_{elec,coul}$	ΔE_{QM}	$\Delta G^{solvation}$	no. of formal charges[b]
1	−58.4	−22.3	−14.6	58.4	0
2	−61.2	−21.2	−14.9	61.3	0
3	−66.6	−18.1	−12.3	62.2	0
4	−64.9	−19.2	−8.7	62.4	0
5	−71.0	−21.2	−15.5	66.0	0
6	−64.6	−31.4	−21.3	74.5	0
7	−67.4	−29.4	−21.5	76.8	0
8	−72.4	−26.1	−20.3	79.2	0
9	−72.1	−24.6	−17.5	81.8	0
10	−77.5	−26.9	−23.4	86.1	0
11	−73.2	−27.5	−25.1	99.2	0
12	−76.2	−24.3	−25.8	102.7	0
13	−80.7	−22.5	−23.8	103.4	0
14	−80.4	−18.8	−16.4	105.6	0
15	−81.2	−22.9	−17.1	100.3	0
16	−75.4	−14.4	−29.3	129.2	0
17	−78.5	−12.2	−30.3	133.0	0
18	−82.6	−10.4	−27.5	133.7	0
19	−82.4	−7.5	−20.2	135.5	0
20	−83.2	−11.9	−21.0	132.5	0
21	−69.1	−56.2	−38.2	69.3	1
22	−71.7	−50.5	−34.3	71.8	1
23	−76.0	−46.2	−29.0	73.7	1
24	−76.2	−60.4	−32.1	64.2	1

[a] All energy values are in kcal·mol^{-1}. [b] Neutral blocking group or positive charge on unblocked N-terminal amino group of inhibitors 21−24. The C-terminal group is neutral; it is $-NH_2$ or −O−Me for inhibitors 1−10 or 11−24, respectively.

CDK2 (PDB code 1KE5) with the inhibitors published by Bramson[38] and Gibson[39] were also available from a previous study.[27]

Minimization. Standard protonation states at neutral pH were used for all ionizable side chains (i.e., neutral His, positively charged Arg and Lys, and negatively charged Asp and Glu) except for one of the two carboxy groups in the Asp catalytic dyad of HIV-1 PR.[24] The net charge of WNV PR, HIV-1 PR, and CDK2 is −10, +7, and +5, respectively. Hydrogen atoms were added to all structures and minimized with the program CHARMM[40] and the CHARMm22 force field[41] (Accelrys Inc.). Partial charges were assigned using the MPEOE method.[42,43] The WNV PR protein/inhibitor complexes were minimized with a two-step protocol. First, the inhibitors were minimized by 200 iterations of the steepest descent algorithm with rigid protein. The second step consisted of 10 000 iterations of the adopted basis Newton−Raphson algorithm to an rms of the gradient of 0.001 kcal mol^{-1} Å^{-1} with flexible protein but using harmonic restraints on all carbon atoms of protein and inhibitor. The value of the force constants was gradually decreased from 20 to 1 kcal mol^{-1} Å^2. In the first minimization the electrostatic energy term was screened by a distance-dependent dielectric function ($\varepsilon(r) = 4r$), and the default nonbonding cutoff of 14 Å was used. In the second minimization the Coulombic energy (constant dielectric of 1.0) was evaluated without truncation. The distance-dependent dielectric in the first minimization and the harmonic restraints on carbon atoms in the second minimization were applied to prevent artificial deviations due to vacuum effects. The second step of the minimization protocol with vacuum dielectric yields optimal QMLIECE model. Several optimization protocols were tested, and it was found that QMLIECE always outperforms LIECE model for WNV PR, irrespective of the protocol (see Supporting Information). Ideally, one should minimize the sum of van der Waals and QM energy contributions. However, optimization using QM, even with linear-scaling techniques, is still computationally too costly for enzyme/ligand complexes.[44] For HIV-1 PR and CDK2 complexes, the inhibitors were optimized by 200 iterations of the

Table 4. Energy Components of the CDK2 Inhibitors[a]

ID	ΔE_{vdW}	$\Delta E_{elec,coul}$	ΔE_{QM}	$\Delta G^{solvation}$
1	−24.5	−10.1	−8.3	36.0
2	−26.5	−10.2	−9.2	30.1
3	−28.2	−9.4	−9.2	32.1
4	−24.2	−11.0	−8.7	35.0
5	−26.7	−10.9	−10.0	34.7
6	−28.5	−8.7	−7.2	34.1
7	−28.9	−9.7	−4.5	28.6
8	−28.8	−8.1	−4.5	35.1
9	−22.3	−9.2	−7.9	29.5
10	−28.3	−10.4	−10.4	39.5
11	−23.2	−10.1	−8.5	35.7
12	−26.6	−11.1	−9.0	36.1
13	−28.5	−11.3	−6.2	33.6
14	−30.6	−8.8	−8.2	39.3
15	−31.2	−13.7	−15.9	46.2
16	−29.7	−10.5	−9.2	37.0
17	−29.3	−10.3	−12.0	40.4
18	−31.2	−9.9	−7.8	37.9
19	−29.9	−10.1	−10.4	39.0
20	−31.7	−1.5	2.1	35.3
21	−34.6	−11.3	−11.1	52.3
22	−31.3	−5.8	−4.2	31.2
23	−32.9	−14.0	−15.9	59.3
24	−39.3	−27.0	−21.8	54.3
25	−43.8	−26.6	−21.0	55.3
26	−43.6	−27.1	−22.9	56.2
27	−46.2	−26.4	−22.6	59.7
28	−47.3	−25.7	−21.0	63.1
29	−46.4	−27.3	−24.1	64.4
30	−44.3	−28.0	−24.4	64.7
31	−46.5	−29.1	−25.2	69.0
32	−48.9	−18.2	−19.3	66.3
33	−51.3	−32.7	−28.9	71.8
34	−51.8	−28.6	−27.4	69.7
35	−46.0	−27.2	−19.6	60.6
36	−42.7	−26.1	−21.9	53.8
37	−38.9	−26.0	−20.3	58.2
38	−40.0	−20.5	−19.2	59.1
39	−40.7	−20.4	−18.8	61.8
40	−42.2	−20.4	−17.8	59.3
41	−40.1	−21.8	−20.7	56.9
42	−39.4	−37.7	−27.4	63.1
43	−38.8	−17.4	−12.1	53.2
44	−39.9	−33.7	−25.8	61.5
45	−38.6	−30.8	−24.5	57.8
46	−41.4	−37.4	−26.7	79.1
47	−40.9	−41.9	−33.4	81.1
48	−41.9	−28.9	−28.7	67.5
49	−45.2	−33.3	−30.8	79.3
50	−50.9	−35.3	−22.4	80.8
51	−50.9	−40.9	−26.7	85.3
52	−44.3	−20.2	−23.3	66.6
53	−39.5	−25.8	−20.7	59.4
54	−42.9	−22.7	−16.1	52.8
55	−36.8	−17.6	−22.0	64.9
56	−42.0	−18.0	−19.3	58.0
57	−41.3	−15.5	−12.0	55.9
58	−41.5	−32.4	−25.6	61.1
59	−39.5	−27.1	−20.1	53.0
60	−39.6	−22.8	−15.5	50.6
61	−36.5	−23.3	−13.5	52.5
62	−40.3	−26.4	−20.1	53.2
63	−37.8	−23.3	−13.1	61.9
64	−41.6	−46.0	−31.8	71.0
65	−40.2	−24.7	−19.4	56.8
66	−40.5	−27.9	−22.2	64.2
67	−42.0	−30.2	−23.4	72.6
68	−42.4	−27.4	−26.2	65.9
69	−44.0	−29.4	−23.4	75.0
70	−47.1	−26.6	−24.8	71.7
71	−46.9	−32.9	−28.0	78.3
72	−47.8	−30.5	−28.2	83.4
73	−43.1	−31.1	−23.9	72.4

[a] All 73 inhibitors are nonpeptidic and devoid of formal charges.[38,39] All energy values are in kcal·mol^{-1}.

Table 5. QMLIECE and LIECE Models[a]

				leave-one-out cross-valid		
	α	β	$\Delta G_{tr,rot,bond}$ (kcal·mol^{-1})	rms error (kcal·mol^{-1})	q^2	rms error on test set[b] (kcal·mol^{-1})
WNV PR (44 Peptidic Inhibitors with $0 \le Q \le 3$)						
$\beta\Delta G_{QM\_elesol} + \Delta G_{tr,rot,bond}$		0.022	−7.6	0.67	0.65	
standard deviation		±0.003	±0.2			
$\beta\Delta G_{MM\_elesol} + \Delta G_{tr,rot,bond}$		0.032	−5.7	0.91	0.35	
standard deviation		±0.006	±0.3			
WNV PR (37 Peptidic Inhibitors with $2 \le Q \le 3$)						
$\beta\Delta G_{QM\_elesol} + \Delta G_{tr,rot,bond}$		0.024	−7.6	0.64	0.70	
standard deviation		±0.004	±0.2			
$\beta\Delta G_{MM\_elesol} + \Delta G_{tr,rot,bond}$		0.048	−5.0	0.84	0.49	
standard deviation		±0.009	±0.4			
HIV-1 PR (24 Peptidic Inhibitors)						
$\alpha\Delta E_{vdW} + \beta\Delta G_{QM\_elesol} + \Delta G_{tr,rot}$	0.350	0.067	8.3	0.64	0.80	1.15
standard deviation	±0.063	±0.025	±2.8			
$\alpha\Delta E_{vdW} + \beta\Delta G_{MM\_elesol} + \Delta G_{tr,rot}$	0.299	0.032	7.9	0.67	0.78	1.30
standard deviation	±0.048	±0.013	±2.8			
CDK2 (Nonpeptidic Inhibitors)						
$\alpha\Delta E_{vdW} + \beta\Delta G_{QM\_elesol}$	0.241	*0.002*[c]		0.99	0.79	
standard deviation	±0.022	*±0.022*[c]				
$\alpha\Delta E_{vdW} + \beta\Delta G_{MM\_elesol}$	0.265	0.029		0.98	0.79	
standard deviation	±0.018	±0.020				

[a] For each set of enzyme/inhibitor complexes the QMLIECE and LIECE models differ only in ΔG_{QM_elesol} and ΔG_{MM_elesol}, respectively (see Methods). [b] The test set was not used to derive the model. It contains five HIV-1 PR inhibitors with K_i values of 0.05, 0.38, 3.2, 437, and 1100 nM. [c] Parameters with leave-one-out standard deviation larger than the average value are statistically not significant and are given in italics.

steepest descent algorithm followed by 10 000 iterations of the adopted basis Newton−Raphson algorithm to an rms of the gradient of 0.001 kcal mol^{-1} Å^{-1} with rigid protein. Because of the predominance of the van der Waals term, which is identical in LIECE and QMLIECE, similar fitting results for these two complexes are obtained by minimizing with rigid protein or harmonically restrained protein (not shown).

Energy Calculation. All QM energy values were calculated on CHARMM-minimized structures. The vacuum interaction energies in QMLIECE were calculated with a divide and conquer approach[9,10] (Figure 1) using MOPAC[45] and the recently developed semiempirical Hamiltonian RM1.[16] The QM energy characterizes the nonclassical charge transfer effect, which is omitted in the fixed-charge model but can be strong if a cation group and an anion group are closed to each other, e.g., the positively charged side chains of Bz-Nle-Lys-Arg-Arg-H and negatively charged sub pockets of WNV PR.[34] van der Waals interactions are fundamentally charge−charge interactions consisting of attractive and repulsive interactions originating from dispersive forces and exchange forces, respectively. The interaction energies from semiempirical QM calculations include the repulsive part of van der Waals interaction energy but ignore the attractive part,[17] which needs highly correlated treatments and large basis sets.[46] Ideally, "pure" electrostatic part of QM interaction energy is needed for linear combination with van der Waals part from molecular mechanics (MM) calculation. However, the QM interaction energy cannot be decomposed as in classical force fields. A linear combination of the QM and MM energy contributions is used in QMLIECE to partially remove the double counting of the repulsive part of van der Waals interaction. In any case, minimized complexes have negligible repulsive interactions.

The van der Waals energy and the MM vacuum Coulombic energy ($\varepsilon(r) = 1$, infinite cutoff) were calculated using CHARMM[40] and the CHARMm[41] force field with the same protocol as in a previous publication.[24]

The electrostatic solvation energy was calculated by the finite-difference Poisson approach using the PBEQ[47] module in CHARMM[40] and a focusing procedure with a final grid spacing of 0.25 Å. The size of the initial grid was determined by considering a layer of at least 22.5 Å around the solute. The dielectric discontinuity interface was delimited by the molecular surface which is spanned by the surface of a rolling probe of 1.4 Å. The ionic strength was set to zero, and the temperature was set to 300 K. Two finite-difference Poisson calculations were performed for each of the three systems (inhibitor, protein, inhibitor/protein complex). The exterior dielectric constant was set to 78.5 and 1.0 for the first and second calculation, respectively, while the solute dielectric constant was set to 1.0, which is consistent with QM energy and parameters of the CHARMm22 force field.

Binding Free Energy. The equations used for the fitting are two-parameter models

$$\Delta G_{bind} = \beta\Delta G_{elesol} + \Delta G_{tr,rot,bond} \quad \text{for WNV PR} \qquad (1)$$

$$\Delta G_{bind} = \alpha\Delta E_{vdW} + \beta\Delta G_{elesol} \quad \text{for CDK2}^{48} \qquad (2)$$

and a three-parameter model

$$\Delta G_{bind} = \alpha\Delta E_{vdW} + \beta\Delta G_{elesol} + \Delta G_{tr,rot} \quad \text{for HIV-1 PR}^{24} \qquad (3)$$

The electrostatic contribution to the binding energy ΔG_{elesol} is the sum of the ligand/protein electrostatic interaction energy in solvent ($\Delta G^{sol}_{pro/lig}$) and the change in solvation energy of ligand and protein upon binding:[49,50]

$$\begin{aligned}\Delta G_{elesol} &= \Delta G^{sol}_{prot/lig} - \Delta G^{solvation}_{prot} - \Delta G^{solvation}_{lig} \\ &= \Delta G^{vacuo}_{prot/lig} + \Delta G^{solvation}_{prot/lig} - \Delta G^{solvation}_{prot} - \Delta G^{solvation}_{lig} \\ &= \Delta G^{vacuo}_{prot/lig} + \Delta G^{solvation}\end{aligned} \qquad (4)$$

For the vacuum electrostatic interaction energy$\Delta G^{vacuo}_{prot/lig}$, QM ($\Delta E_{QM}$ in Table 2–4) and MM ($\Delta E_{elec,coul}$) calculations were used in QMLIECE and LIECE, respectively. Note that ΔE_{QM} could be further decomposed into electrostatic and explicit polarization energy terms,[31,32] but such decomposition would require additional fitting parameters. The finite-difference Poisson approach was used to calculate the solvation energy changes upon binding ($\Delta G^{solvation}$). Details are given in the Supporting Information.

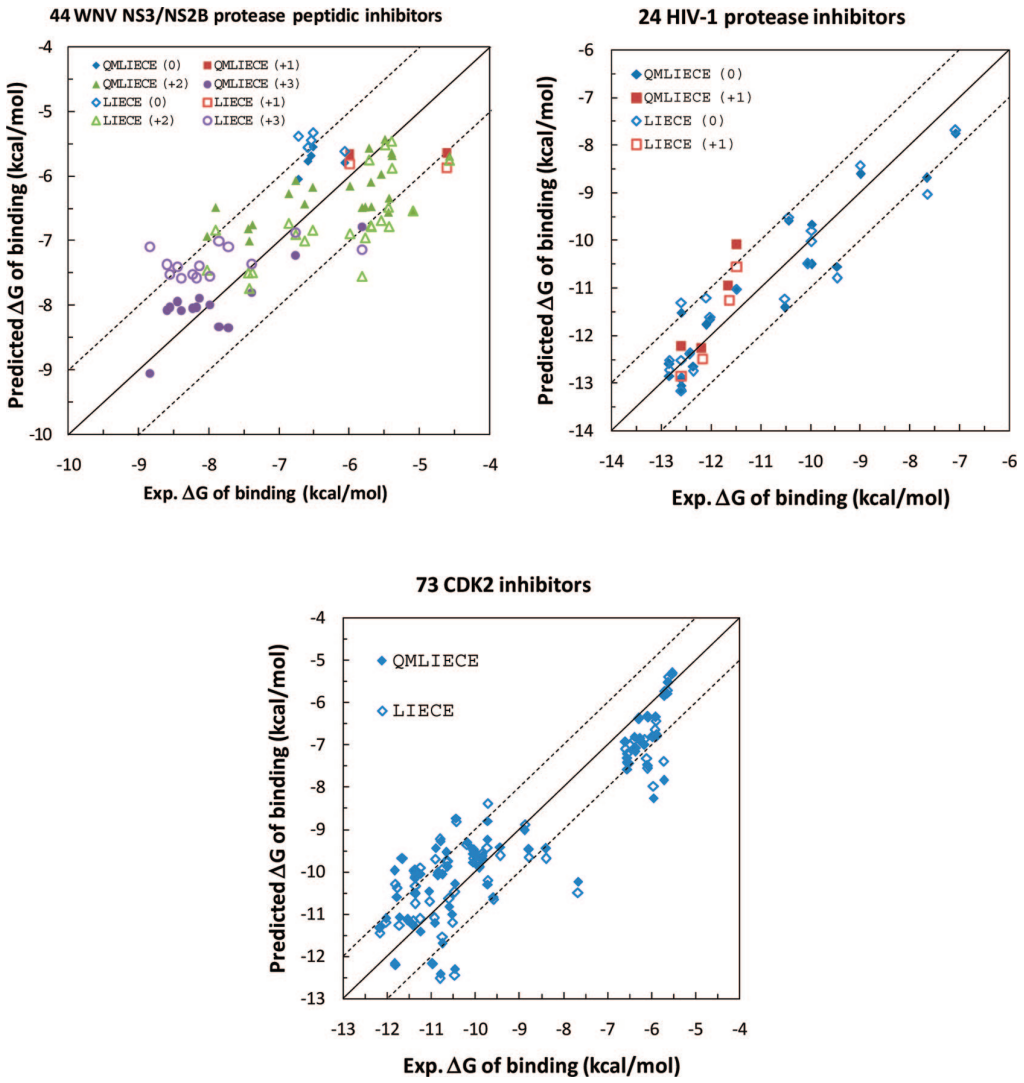


Figure 2. Comparison of the calculated (QMLIECE filled symbols, LIECE empty symbols) versus experimental binding free energies for 44 WNV PR[34–36] (top left), 24 HIV-1 PR[37] (top right), and 73 CDK2[38,39] (bottom) inhibitors. The experimental data are fitted with two-parameter models for WNV PR (eq 1), three-parameter models for HIV-1 PR (eq 3), and two-parameter models for CDK2 (eq 2). Digit in parentheses is the total charge of the inhibitor.

The term ΔE_{vdW} is the ligand/protein van der Waals interaction energy. Since the semiempirical QM calculation does not take into account the attractive part of the van der Waals energy, the van der Waals interaction energy of the force field is still used in QMLIECE.

The constant term $\Delta G_{tr,rot,bond}$ accounts for the loss of translational and rotational degrees of freedom upon binding and the energy of formation of the covalent bond for the 44 aldehyde inhibitors of WNV PR. The entropic penalty due to loss of translational and rotational degrees of freedom ($\Delta G_{tr,rot}$) is unfavorable and therefore positive, but its sum with the covalent bond energy can also be negative.

For WNV PR (eq 1), ΔE_{vdW} was neglected because the statistical significance of the fitting deteriorates (see Supporting Information). The same is observed upon addition of $\Delta G_{tr,rot}$ in the CDK2 models, which is consistent with the significantly smaller flexibility of the nonpeptidic inhibitors of CDK2 than the peptidic inhibitors of WNV PR and HIV-1 PR.

3. Results and Discussion

The energy values and the parameters obtained by least-squares fitting are given in Tables 2–4 and Table 5, respectively, while the correlation between LIECE/QMLIECE binding energies and experimental values is shown in Figure 2.

WNV PR. The two-parameter QMLIECE model yields a leave-one-out rms of the error of 0.67 kcal·mol^{-1} and cross-validated q^2 of 0.65. These results are significantly better than those obtained by LIECE (rms error of 0.91 kcal·mol^{-1} and cross-validated q^2 of 0.35). As an additional test, the LIECE and QMLIECE models were applied to a nonpeptidic inhibitor, discovered recently in our group (Ekonomiuk et al., unpublished results), which was not used to derive the model. The LIECE and QMLIECE binding affinity are −5.5 and −8.6 kcal·mol^{-1}, respectively, while the experimentally measured binding affinity is −7.2 kcal·mol^{-1}. Since this inhibitor does not bind covalently to the protein, the calculated binding free energy should be more favorable than the measured value because the LIECE and

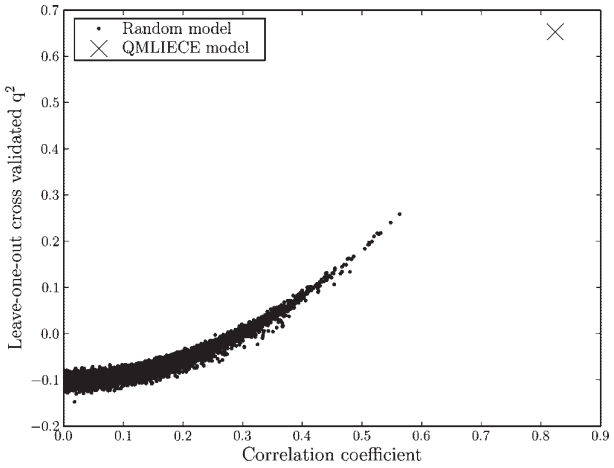


Figure 3. Statistical test to assess the predictive power of the QMLIECE two-parameter model for WNV PR (cross) by comparison with 10 000 models obtained by randomizing the activity values (dots). The QMLIECE model is clearly separated, which indicates that QMLIECE not only better fits the data than the random models but also has a better predictive ability. In other words, the plot shows that the QMLIECE model does not suffer from chance correlation.

QMLIECE models for WNV PR were derived from 44 peptidic inhibitors covalently bound to the protein. Therefore, the QMLIECE value is more reliable than the LIECE one.

A statistical test based on the randomization of the data points was used to analyze an eventual chance correlation between the QMLIECE model and the data points.[25,51] The binding free energies of the 44 peptidic inhibitors[34–36] were randomly chosen from uniformly distributed values in the same range as the experimental values (i.e., from −8.84 to −4.58 kcal·mol^{-1}), and the multiplicative parameters of ΔG_{elesol} and $\Delta G_{tr,rot,bond}$ constant term were determined by fitting to random "data points". The rationale behind this test is that the statistical significance of the real model is poor if there is a significant correlation between the descriptors and the randomized data points. The randomization and fitting were repeated 10 000 times, and Figure 3 shows the cross-validated q^2 (obtained by the leave-one-out procedure) plotted versus the correlation coefficient. The QMLIECE model with the two parameters fitted to the real data points is located at the top right corner and is significantly separated from the models generated by the randomization of the binding free energies. This separation provides further evidence that the QMLIECE two-parameter model not only fits the experimental data but also has very good predictive ability, i.e., chance correlation is not present. To further assess the significance of the QMLIECE model, the same statistical test was performed on two naive models suggested by an anonymous reviewer: a combination of LIECE and a binary descriptor for distinguishing inhibitors with charge +2 from all others and a simple five-parameter model based only on binary descriptors for the number of positive charges in the inhibitors (Supporting Information). Interestingly, the risk of chance correlation increases with increasing number of fitting parameters and decreasing physical soundness. In other words, for the QMLICE model only, there is a genuine correlation between descriptors and data points.

HIV-1 PR. The three-parameter QMLIECE model yields an rms of the error of 0.64 kcal·mol^{-1} and a cross-validated q^2 of 0.80. These results are similar to those obtained by LIECE (rms error of 0.67 kcal·mol^{-1} and a cross-validated q^2 of 0.78). (Note that the rms error of 0.67 kcal·mol^{-1} is slightly smaller than in ref 24, where it was 0.77 kcal·mol^{-1}, because of the different minimization protocol.) The predictive ability of the LIECE and QMLIECE approach was further tested on a set of five inhibitors available from a previous work[52] and not used to derive the models. Their PDB identifiers and K_i values are the following: 1HVR, K_i = 0.05 nM, K_i = 0.38 nM; 1HTG, K_i = 3.2 nM;

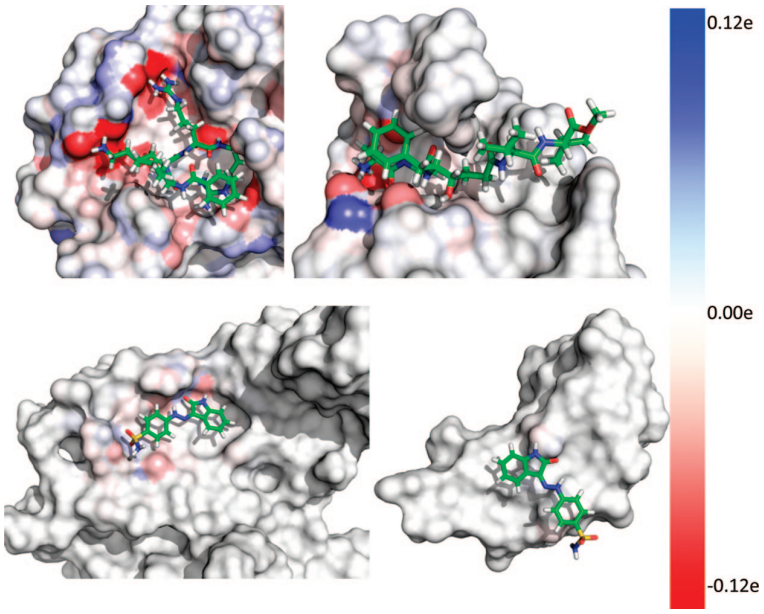


Figure 4. Polarization of protein atoms due to inhibitor binding: WNV PR in complex with its inhibitor 1 (top left), HIV-1 PR in complex with its inhibitor 21 (top right, only one monomer of the C_2-symmetric structure of the HIV-1 PR homodimer is shown), and CDK2 in complex with its inhibitor 24 (α-helical domain bottom left and β-sheet domain bottom right). The polarized charges were calculated by subtracting SCF atomic charges before binding from that after binding, using the divide and conquer protocol.[9] The protein surfaces were rendered with the blue-white-red spectrum according to polarized charges of atoms. The blue on the surface denotes atomic partial charges that become more positive upon binding, while red means more negative atomic charges upon binding, and white indicates atomic charges that do not change upon binding.

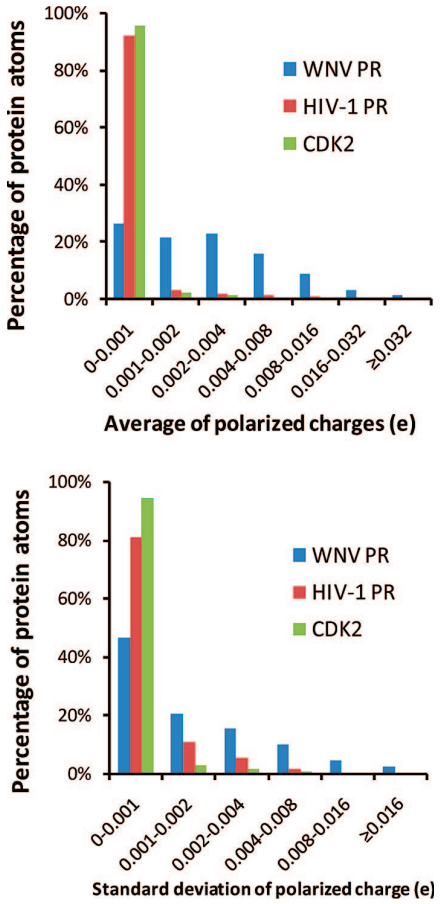


Figure 5. Distribution of average (top) and standard deviation (bottom) of individual polarized charges of proteins upon binding, calculated over all inhibitors (44, 24, and 73 inhibitors for WNV PR, HIV-1 PR, and CDK2, respectively). The polarized charges were calculated by subtracting SCF atomic charges before binding from that after binding, using the divide and conquer protocol.[9]

1HBV, $K_i = 437$ nM; 1HVS, 5HVP, $K_i = 1100$ nM.[53–57] The five inhibitors were minimized in the HIV-1 PR conformation from the 1HVR complex because of its highest resolution (1.8 Å). The error rms for the five inhibitors of the test set is 1.30 and 1.15 kcal·mol^{-1} for the LIECE and QMLIECE models, respectively. This comparison indicates that for the 24 mainly neutral inhibitors of HIV-1 PR the QMLIECE model is only slightly more predictive than the LIECE model.

CDK2. The two-parameter model of QMLIECE yields an rms of the error of 0.99 kcal·mol^{-1} and a cross-validated q^2 of 0.79. This accuracy is essentially identical to the one of the LIECE model. Moreover, the electrostatic parameter of the QMLIECE model is smaller than the standard deviation obtained by the leave-one-out procedure, which indicates that the QMLIECE model of CDK2 is not robust.

Applicability of QM. It is important to clarify under which circumstance it is necessary to use QM for calculating electrostatic energy contribution in linear interaction energy models. The advantage of QM compared with MM is that QM allows the evaluation of charge-transfer effects by self-consistent field (SCF) calculation. Upon inhibitor binding, the amount of polarization of WNV PR is larger than HIV-1 PR and much larger than CDK2 (Figure 4). As a matter of fact, for the complexes of HIV-1 PR and CDK2 the charge−charge interactions are relatively similar and small (Figure 4). Furthermore, more than 90% of their atoms are not significantly polarized (less than 0.001e) upon inhibitor binding (Figure 5). Therefore, the absolute errors originating from polarization are small for these two enzyme/inhibitor complexes, and can be rectified by the regression parameters without leading to poor fitting. On the other hand, the 44 peptidic inhibitors of WNV have between zero and three positively charged side chains resulting in a large variability of polarized charges; as a consequence, the energies calculated by MM are significantly different from QM energies because only the latter takes charge polarization effects into account. This explains the better predictive ability of QMLIECE than LIECE for the 44 inhibitors of WNV PR.

An additional test was performed to separate the effect of total charge from charge variability. The inhibitors of WNV PR with zero or one formal charge (7 of the 44 inhibitors) were removed from the fitting data of two-parameter model of LIECE. The variability of polarized charges, therefore, becomes smaller for this subset, while the average value of polarized charges becomes even larger. By application of leave-one-out cross-validation to the reduced set of data (37 inhibitors), it is found that QMLIECE does not change significantly whereas q^2 of LIECE improves from 0.35 to 0.49, Moreover, in the LIECE model generated using only 37 inhibitors with $2 \leq Q \leq 3$ the parameter of ΔG_{MM_elesol} increases from 0.032 to 0.048, and the constant term $\Delta G_{tr,rot,bond}$ changes from −5.7 to −5.0 kcal·mol^{-1} (Table 5). These results indicate that the weight of the electrostatic contribution in the LIECE regression model increases by reducing the formal charge variability of inhibitors despite the larger average total charge. In other words, the neglect of polarization in LIECE results in acceptable predictive ability for binding affinities of inhibitors with two or three formal charges.

Therefore, if the charge−charge interactions between inhibitors and protein are similar, even though the absolute values of them are large, the fixed charge model in the force field method can attain reasonable results for the evaluation of electrostatic energies. Otherwise, QM is needed to more accurately evaluate the variable influence of polarization effects on electrostatic interactions.

Computational Requirements. The time for the QM calculation is linearly related to the number of residues. For WNV PR (187 residues), the QM calculation needs about 40 min on an Opteron 252 (2.6 GHz). The total time required by QMLIECE is about 1 h for each inhibitor. The finite-difference Poisson and QM calculations require about 900 and 50 MB memory, respectively.

4. Conclusions

Previously the computationally expensive sampling by MD in the linear interaction energy model had been substituted with a simple energy minimization and continuum electrostatics calculation (LIECE).[24] QMLIECE is a further development of LIECE, in which the force field based electrostatic part of the inhibitor/protein interaction energy is replaced by the corresponding contribution evaluated by a QM calculation at the semiempirical RM1[16] level with a linear-scaling method. LIECE and QMLIECE models are assessed on three classes of inhibitor/enzyme complexes: 44 inhibitors of a flaviviral serine protease, 24 inhibitors of a retroviral aspartic protease, and 73 inhibitors of the CDK2 serine/threonine protein kinase. Only for the 44 inhibitors of the serine protease, which have between zero and three positive charges, did QMLIECE show a significant improvement compared to LIECE. However, for the subset of 37 of the 44 inhibitors with two or three positive charges LIECE

was more predictive than for the full set of 44 inhibitors but still not as robust as QMLIECE. Therefore, the comparison of LIECE and QMLIECE indicates that the use of QM is necessary when complexes with different inhibitors have significantly diverse charge−charge interactions, i.e., a large variability of polarized charges of protein atoms upon binding different inhibitors.

Acknowledgment. This work was supported in part by the Hartmann-Müller Foundation. Calculations were performed on the Matterhorn Beowulf cluster.

Supporting Information Available: Tables of different optimization protocols for QMLIECE and LIECE, including CHARMM commands, their cross-validated q^2, three-parameter LIECE and QMLIECE models, and statistical tests on simple models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–1818.
(2) Warshel, A.; Karplus, M. Calculation of ground and excited-state potential surfaces of conjugated molecules. 1. Formulation and parametrization. *J. Am. Chem. Soc.* **1972**, *94*, 5612–5625.
(3) Field, M. J.; Bash, P. A.; Karplus, M. A combined quantum-mechanical and molecular mechanical potential for molecular-dynamics simulations. *J. Comput. Chem.* **1990**, *11*, 700–733.
(4) Masgrau, L.; Roujeinikova, A.; Johannissen, L. O.; Hothi, P.; Basran, J.; Ranaghan, K. E.; Mulholland, A. J.; Sutcliffe, M. J.; Scrutton, N. S.; Leys, D. Atomic description of an enzyme reaction dominated by proton tunneling. *Science* **2006**, *312*, 237–241.
(5) Pu, J. Z.; Gao, J. L.; Truhlar, D. G. Multidimensional tunneling, recrossing, and the transmission coefficient for enzymatic reactions. *Chem. Rev.* **2006**, *106*, 3140–3169.
(6) Gao, J. L.; Ma, S. H.; Major, D. T.; Nam, K.; Pu, J. Z.; Truhlar, D. G. Mechanisms and free energies of enzymatic reactions. *Chem. Rev.* **2006**, *106*, 3188–3209.
(7) Senn, H. M.; Thiel, W. QM/MM studies of enzymes. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182–187.
(8) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. Molecular tailoring approach for simulation of electrostatic properties. *J. Phys. Chem.* **1994**, *98*, 9165–9169.
(9) Dixon, S. L.; Merz, K. M. Semiempirical molecular orbital calculations with linear system size scaling. *J. Chem. Phys.* **1996**, *104*, 6643–6649.
(10) Zhang, D. W.; Zhang, J. Z. H. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein−molecule interaction energy. *J. Chem. Phys.* **2003**, *119*, 3599–3605.
(11) Raha, K.; Merz, K. M. Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein−ligand complexes. *J. Med. Chem.* **2005**, *48*, 4558–4575.
(12) Cavalli, A.; Carloni, P.; Recanatini, M. Target-related applications of first principles quantum chemical methods in drug design. *Chem. Rev.* **2006**, *106*, 3497–3519.
(13) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; WollaCott, A. M.; Westerhoff, L. M.; Merz, K. M. The role of quantum mechanics in structure-based drug design. *Drug Discovery Today* **2007**, *12*, 725–731.
(14) Vondrasek, J.; Bendova, L.; Klusak, V.; Hobza, P. Unexpectedly strong energy stabilization inside the hydrophobic core of small protein rubredoxin mediated by aromatic residues: Correlated ab initio quantum chemical calculations. *J. Am. Chem. Soc.* **2005**, *127*, 2615–2619.
(15) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.
(16) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RMI: a reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27*, 1101–1111.
(17) Raha, K.; Merz, K. M. A quantum mechanics-based scoring function: study of zinc ion-mediated ligand binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.
(18) Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *8*, 365–370.
(19) Joseph-McCarthy, D.; Baber, J. C.; Feyfant, E.; Thompson, D. C.; Humblet, C. Lead optimization via high-throughput molecular docking. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 264–274.
(20) Giese, T. J.; York, D. M. High-level ab initio methods for calculation of potential energy surfaces of van der Waals complexes. *Int. J. Quantum Chem.* **2004**, *98*, 388–408.
(21) Barratt, E.; Bingham, R. J.; Warner, D. J.; Laughton, C. A.; Phillips, S. E. V.; Homans, S. W. van der Waals interactions dominate ligand−protein association in a protein binding site occluded from solvent water. *J. Am. Chem. Soc.* **2005**, *127*, 11827–11834.
(22) Johnson, E. R.; Becke, A. D. van der Waals interactions from the exchange hole dipole moment: application to bio-organic benchmark systems. *Chem. Phys. Lett.* **2006**, *432*, 600–603.
(23) Gonzalez-Diaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. *J. Proteome Res.* **2007**, *6*, 904–908.
(24) Huang, D.; Caflisch, A. Efficient evaluation of binding free energy using continuum electrostatics solvation. *J. Med. Chem.* **2004**, *47*, 5791–5797.
(25) Huang, D. Z.; Luthi, U.; Kolb, P.; Cecchini, M.; Barberis, A.; Caflisch, A. In silico discovery of beta-secretase inhibitors. *J. Am. Chem. Soc.* **2006**, *128*, 5436–5443.
(26) Huang, D. Z.; Luthi, U.; Kolb, P.; Edler, K.; Cecchini, M.; Audetat, S.; Barberis, A.; Caflisch, A. Discovery of cell-permeable non-peptide inhibitors of beta-secretase by high-throughput docking and continuum electrostatics calculations. *J. Med. Chem.* **2005**, *48*, 5108–5111.
(27) Kolb, P.; Huang, D.; Dey, F.; Caflisch, A. Discovery of kinase inhibitors by high-throughput docking and scoring based on a transferable linear interaction energy model. *J. Med. Chem.* **2008**, *51*, 1179–1188.
(28) Hansson, T.; Aqvist, J. Estimation of binding free energies for HIV proteinase inhibitors by molecular dynamics simulations. *Protein Eng.* **1995**, *8*, 1137–1144.
(29) Aqvist, J.; Medina, C.; Samuelsson, J. E. New method for predicting binding-affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385–391.
(30) Aqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. Ligand binding affinities from MD simulations. *Acc. Chem. Res.* **2002**, *35*, 358–365.
(31) Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. L. Importance of substrate and cofactor polarization in the active site of dihydrofolate reductase. *J. Mol. Biol.* **2003**, *327*, 549–560.
(32) Hensen, C.; Hermann, J. C.; Nam, K. H.; Ma, S. H.; Gao, J. L.; Holtje, H. D. A combined QM/MM approach to protein−ligand interactions: polarization effects of the HIV-1 protease on selected high affinity inhibitors. *J. Med. Chem.* **2004**, *47*, 6673–6680.
(33) Erbel, P.; Schiering, N.; D'Arcy, A.; Renatus, M.; Kroemer, M.; Lim, S. P.; Yin, Z.; Keller, T. H.; Vasudevan, S. G.; Hommel, U. Structural basis for the activation of flaviviral NS3 proteases from dengue and West Nile virus. *Nat. Struct. Mol. Biol.* **2006**, *13*, 372–373.
(34) Knox, J. E.; Ma, N. L.; Yin, Z.; Patel, S. J.; Wang, W. L.; Chan, W. L.; Rao, K. R. R.; Wang, G.; Ngew, X.; Patel, V.; Beer, D.; Lim, S. P.; Vasudevan, S. G.; Keller, T. H. Peptide inhibitors of West Nile NS3 protease: SAR study of tetrapeptide aldehyde inhibitors. *J. Med. Chem.* **2006**, *49*, 6585–6590.
(35) Yin, Z.; Patel, S. J.; Wang, W. L.; Chan, W. L.; Rao, K. R. R.; Wang, G.; Ngew, X.; Patel, V.; Beer, D.; Knox, J. E.; Ma, N. L.; Ehrhardt, C.; Lim, S. P.; Vasudevan, S. G.; Keller, T. H. Peptide inhibitors of dengue virus NS3 protease. Part 2: SAR study of tetrapeptide aldehyde inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 40–43.
(36) Yin, Z.; Patel, S. J.; Wang, W. L.; Wang, G.; Chan, W. L.; Rao, K. R. R.; Alam, J.; Jeyaraj, D. A.; Ngew, X.; Patel, V.; Beer, D.; Lim, S. P.; Vasudevan, S. G.; Keller, T. H. Peptide inhibitors of dengue virus NS3 protease. Part 1: Warhead. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 36–39.
(37) Dreyer, G. B.; Lambert, D. M.; Meek, T. D.; Carr, T. J.; Tomaszek, T. A.; Fernandez, A. V.; Bartus, H.; Cacciavillani, E.; Hassell, A. M.; Minnich, M.; Petteway, S. R.; Metcalf, B. W.; Lewis, M. Hydroxyethylene isostere inhibitors of human immunodeficiency virus-1 protease - structure activity analysis using enzyme-kinetics, X-ray crystallography, and infected T-cell assays. *Biochemistry* **1992**, *31*, 6646–6659.
(38) Bramson, H. N.; Corona, J.; Davis, S. T.; Dickerson, S. H.; Edelstein, M.; Frye, S. V.; Gampe, R. T.; Harris, P. A.; Hassell, A.; Holmes, W. D.; Hunter, R. N.; Lackey, K. E.; Lovejoy, B.; Luzzio, M. J.; Montana, V.; Rocque, W. J.; Rusnak, D.; Shewchuk, L.; Veal, J. M.; Walker, D. H.; Kuyper, L. F. Oxindole-based inhibitors of cyclin-dependent kinase 2 (CDK2): design, synthesis, enzymatic activities, and X-ray crystallographic analysis. *J. Med. Chem.* **2001**, *44*, 4339–4358.
(39) Gibson, A. E.; Arris, C. E.; Bentley, J.; Boyle, F. T.; Curtin, N. J.; Davies, T. G.; Endicott, J. A.; Golding, B. T.; Grant, S.; Griffin, R. J.; Jewsbury, P.; Johnson, L. N.; Mesguiche, V.; Newell, D. R.; Noble, M. E. M.; Tucker, J. A.; Whitfield, H. J. Probing the ATP ribose-binding domain of cyclin-dependent kinases 1 and 2 with O6 substituted guanine derivatives. *J. Med. Chem.* **2002**, *45*, 3381–3393.
(40) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Charmm, a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*

1983, *4*, 187–217.
(41) Momany, F. A.; Rone, R. Validation of the general-purpose Quanta 3.2/CHARMm force-field. *J. Comput. Chem.* **1992**, *13*, 888–900.
(42) No, K. T.; Grant, J. A.; Jhon, M. S.; Scheraga, H. A. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 2. Application to ionic and aromatic-molecules as models for polypeptides. *J. Phys. Chem.* **1990**, *94*, 4740–4746.
(43) No, K. T.; Grant, J. A.; Scheraga, H. A. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 1. Application to neutral molecules as models for polypeptides. *J. Phys. Chem.* **1990**, *94*, 4732–4739.
(44) Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. Fully quantum mechanical energy optimization for protein−ligand structure. *J. Comput. Chem.* **2004**, *25*, 1431–1437.
(45) Stewart, J. J. P. Optimization of parameters for semiempirical methods. 1. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.
(46) Chalasinski, G.; Szczesniak, M. M. State of the art and challenges of the ab initio theory of intermolecular interactions. *Chem. Rev.* **2000**, *100*, 4227–4252.
(47) Im, W.; Beglov, D.; Roux, B. Continuum solvation model: computation of electrostatic forces from numerical solutions to the Poisson−Boltzmann equation. *Comput. Phys. Commun.* **1998**, *111*, 59–75.
(48) Kolb, P.; Huang, D.; Dey, F.; Caflisch, A. Discovery of kinase inhibitors by high-throughput docking and scoring based on a transferable linear interaction energy model. *J. Med. Chem.* **2008**, *51*, 1179–1188.
(49) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 88–105.
(50) Majeux, N.; Scarsi, M.; Caflisch, A. Efficient electrostatic solvation model for protein-fragment docking. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 256–268.
(51) So, S. S.; Karplus, M. Genetic neural networks for quantitative structure−activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/GABA(A) receptors. *J. Med. Chem.* **1996**, *39*, 5246–5256.
(52) Cecchini, M.; Kolb, P.; Majeux, N.; Caflisch, A. Automated docking of highly flexible ligands by genetic algorithms: a critical assessment. *J. Comput. Chem.* **2004**, *25*, 412–422.
(53) Fitzgerald, P. M. D.; Mckeever, B. M.; Vanmiddlesworth, J. F.; Springer, J. P.; Heimbach, J. C.; Leu, C. T.; Herber, W. K.; Dixon, R. A. F.; Darke, P. L. Crystallographic analysis of a complex between human-immunodeficiency-virus type-1 protease and acetyl-pepstatin at 2.0-Å resolution. *J. Biol. Chem.* **1990**, *265*, 14209–14219.
(54) Jhoti, H.; Singh, O. M. P.; Weir, M. P.; Cooke, R.; Murrayrust, P.; Wonacott, A. X-ray crystallographic studies of a series of penicillin-derived asymmetric inhibitors of HIV-1 protease. *Biochemistry* **1994**, *33*, 8417–8427.
(55) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C. H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Ericksonviitanen, S. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* **1994**, *263*, 380–384.
(56) Baldwin, E. T.; Bhat, T. N.; Liu, B. S.; Pattabiraman, N.; Erickson, J. W. Structural basis of drug-resistance for the V82a mutant of HIV-1 proteinase. *Nat. Struct. Biol.* **1995**, *2*, 244–249.
(57) Hoog, S. S.; Zhao, B. G.; Winborne, E.; Fisher, S.; Green, D. W.; Desjarlais, R. L.; Newlander, K. A.; Callahan, J. F.; Moore, M. L.; Huffman, W. F.; Abdelmeguid, S. S. A check on rational drug design. Crystal-structure of a complex of human-immunodeficiency-virus type-1 protease with a novel gamma-turn mimetic inhibitor. *J. Med. Chem.* **1995**, *38*, 3246–3252.

JM800242Q

Supporting Information

Is quantum mechanics necessary for predicting binding free energy?

*Ting Zhou, Danzhi Huang and Amedeo Caflisch**
Department of Biochemistry, University of Zürich,
Winterthurerstrasse 190, CH-8057
Zürich, Switzerland

1 Robustness upon choice of minimization protocol

Protocol	Nonbond ID	Minimization parameter ID	Leave-one-out cv q2	
			QMLIECE [a]	LIECE
1	1	1	0.55	0.30
2	3	1	0.59	0.32
3	4	2	0.53	0.06
4	5	2	0.51	0.06
5	2	1	0.46	0.12
6	5	1	0.57	0.39
7	3	3	0.48	0.31
8	3	4	0.46	0.31
9	3	5	0.45	0.32

Nonbond ID	CHARMM nonbond-specification [b]
1	NBONDS NBXMOD 5 GROUP SWITCH CDIE VDW VSWI EXTEND GRAD QUAD CUTNB 180 WMIN 1.5 EPS 1.0
2	NBONDS NBXMOD 5 ATOM SWITCH VATOM VSWITCHED CUTNB 15.0 CTONNB 11 CTOFNB 14 EPS 1.0 E14FAC 0.5 WMIN 1.5 CDIE
3	NBONDS NBXMOD 5 ATOM SWITCH VATOM VSWITCHED CUTNB 180 EPS 1.0 E14FAC 0.5 WMIN 1.5 CDIE
4	NBONDS ATOM FSHIFT CDIE VDW VSHIFT CUTNB 180 WMIN 1.5 EPS 1.0
5	NBONDS ATOM FSWITCH CDIE VDW VSHIFT CUTNB 180 WMIN 1.5 EPS 1.0

Minimization parameter ID	CHARMM minimization commands [c]
1	cons harm force 20 sele (type C*) end mini sd nstep 200 cons harm force 10 sele (type C*) end mini abnr nstep 200 cons harm force 8 sele (type C*) end mini abnr nstep 200 cons harm force 6 sele (type C*) end mini abnr nstep 200 cons harm force 4 sele (type C*) end mini abnr nstep 200 cons harm force 2 sele (type C*) end mini abnr nstep 200 cons harm force 1 sele (type C*) end mini abnr nstep 10000 tolgrd 0.01
2	cons harm force 20 sele (protein and (type C*) or (type O*) or (type N*)) end mini sd nstep 200 cons harm force 10 sele (protein and (type C*) or (type O*) or (type N*)) end mini abnr nstep 200 cons harm force 8 sele (protein and (type C*) or (type O*) or (type N*)) end mini abnr nstep 200 cons harm force 6 sele (protein and (type C*) or (type O*) or (type N*)) end mini abnr nstep 200 cons harm force 4 sele (protein and (type C*) or (type O*) or (type N*)) end mini abnr nstep 200 cons harm force 2 sele (protein and (type C*) or (type O*) or (type N*)) end mini abnr nstep 200 cons harm force 1 sele (protein and (type C*) or (type O*) or (type N*)) end mini abnr nstep 10000 tolgrd 0.01
3	cons harm force 20 sele (type C*) end mini sd nstep 200 cons harm force 10 sele (type C*) end mini abnr nstep 200 cons harm force 8 sele (type C*) end mini abnr nstep 200 cons harm force 6 sele (type C*) end mini abnr nstep 200 cons harm force 4 sele (type C*) end mini abnr nstep 200 cons harm force 2 sele (type C*) end mini abnr nstep 200 cons harm force 1.1 sele (type C*) end mini abnr nstep 10000 tolgrd 0.01
4	cons harm force 20 sele (type C*) end mini sd nstep 200 cons harm force 10 sele (type C*) end mini abnr nstep 200 cons harm force 8 sele (type C*) end mini abnr nstep 200 cons harm force 6 sele (type C*) end mini abnr nstep 200 cons harm force 4 sele (type C*) end mini abnr nstep 200 cons harm force 2 sele (type C*) end mini abnr nstep 200 cons harm force 1.2 sele (type C*) end mini abnr nstep 10000 tolgrd 0.01
5	cons harm force 20 sele (type C*) end mini sd nstep 200 cons harm force 10 sele (type C*) end mini abnr nstep 200 cons harm force 8 sele (type C*) end mini abnr nstep 200 cons harm force 6 sele (type C*) end mini abnr nstep 200 cons harm force 4 sele (type C*) end mini abnr nstep 200 cons harm force 2 sele (type C*) end mini abnr nstep 200 cons harm force 1.3 sele (type C*) end mini abnr nstep 10000 tolgrd 0.01

Table 1. Influence of minimization protocol on fitting results for 44 peptidic inhibitors of WNV PR. [a] These leave-one-out q^2 of QMLIECE were calculated using semiempirical Hamiltonian PM3. We finally chose RM1 Hamiltonian because we found that RM1 is more sophisticated than PM3, and could attain higher fitting qualities. [b] The explanation of charmm nonbond-specification can be found at http://www.charmm.org/html/documentation/c34b1/nbonds.html . [c] The explanation of charmm minimization commands can be found at http://www.charmm.org/html/documentation/c34b1/minimiz.html .

2 Three-parameter LIECE and QMLIECE models

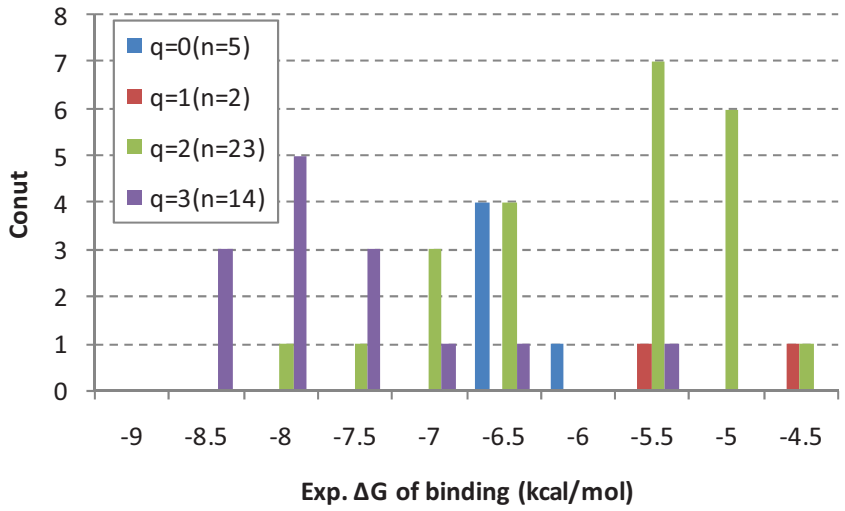
	α	β	$\Delta G_{tr,rot,bond}$ (kcal·mol^{-1})	Leave-one-out cross-valid rms error (kcal·mol^{-1})	q^2
WNV PR (44 peptidic inhibitors)					
$\alpha\Delta E_{vdW}+\beta\Delta G_{QM\_elesol}+\Delta G_{tr,rot,bond}$	0.036	0.026	-6.1	0.65	0.66
Standard deviation	±0.031	±0.005	±1.3		
$\alpha\Delta E_{vdW}+\beta\Delta G_{MM\_elesol}+\Delta G_{tr,rot,bond}$	*-0.012* [a]	0.030	-6.3	0.91	0.32
Standard deviation	*±0.029*	±0.008	±1.6		
CDK2 (73 non-peptidic inhibitors)					
$\alpha\Delta E_{vdW}+\beta\Delta G_{QM\_elesol}+\Delta G_{tr,rot}$	0.264	*-0.0015* [a]	1.0	0.97	0.79
Standard deviation	±0.026	±0.023	±0.6		
$\alpha\Delta E_{vdW}+\beta\Delta G_{MM\_elesol}+\Delta G_{tr,rot}$	0.283	0.023	0.9	0.96	0.79
Standard deviation	±0.021	±0.020	±0.6		

Table 2. Three-parameter LIECE and QMLIECE models. [a] Parameters with leave-one-out standard deviation larger than the average value are statistically not significant and are given in italics.

3 QMLIECE vs. simple models (WNV PR)

3.1 Histogram of ΔG_{bind} values

The 44 inhibitors can be separated according to number of positively charged groups as in the histograms enclosed. By inspection of the histograms one could speculate that the activity is related simply to the number of charges except for the compounds with q=2 (and q=1 which are only two) which seem shifted to less negative values.



3.2 Statistical tests

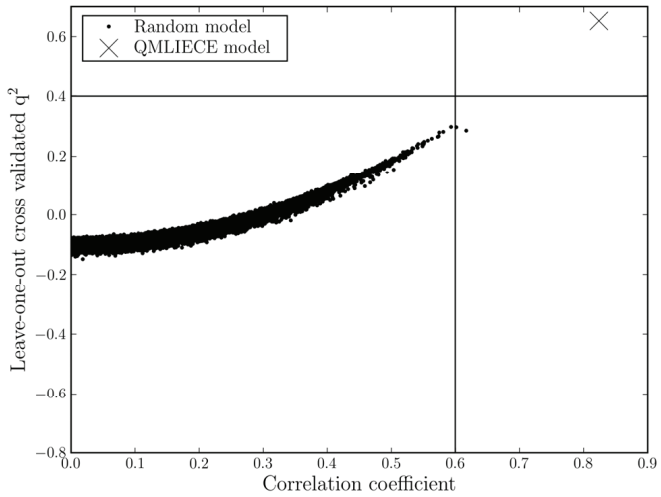
In response to an anonymous reviewer we compare the probability of chance correlation of QMLIECE and simple models derived using the total charge.
The QMLIECE model is the two-parameter model as in Equation (1) in the main text.
The LIECE and binary (LB) is a three-parameter model

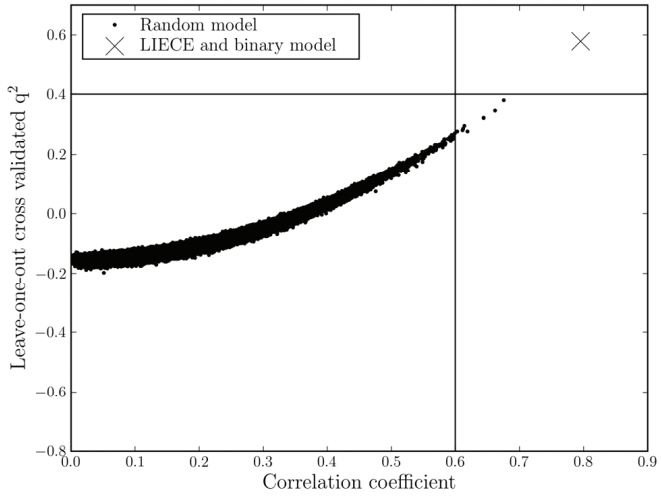
$$\Delta G_{bind} = 0.0318\Delta G_{MM\_elesol} + 1.1307Q_2 - 6.3154$$

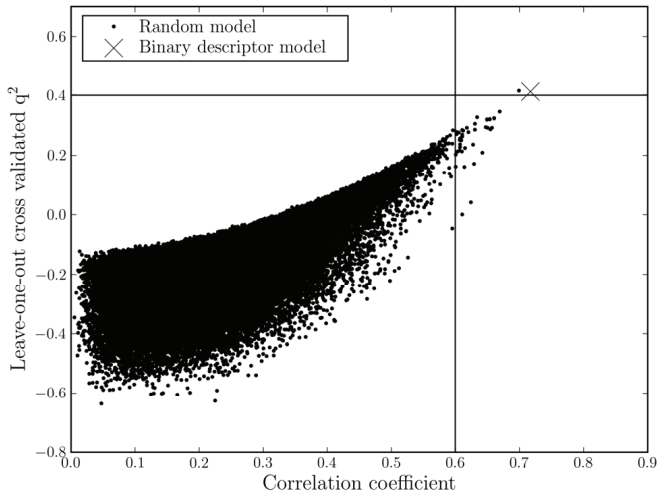
where ΔG_{MM_elesol} is the electrostatic contribution of binding free energy calculated by force field method, and Q_2 is a binary descriptor described below..
The binary descriptor (BD) is a five-parameter model

$$\Delta G_{bind} = 1.141Q_0 + 2.325Q_1 + 1.451Q_2 - 0.350Q_3 - 7.625$$

where "Q_n" is a binary descriptor, equals to one for ligands with charge of *n* and zero for all others.





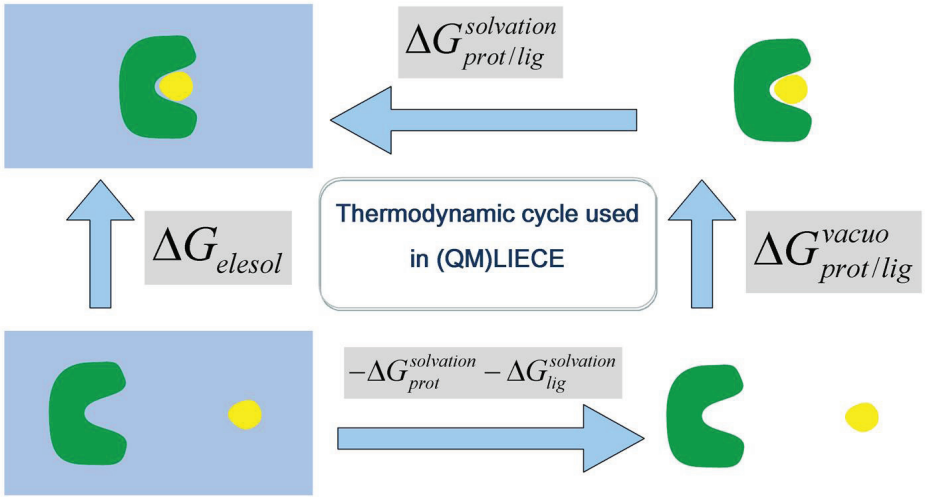


These statistical tests evaluate the predictive power of QMLIECE and two simple models. For each of the three models a total of 100000 models are generated by random guessing of the binding affinities (see main text for details).

The rationale behind this statistical test is that the significance of the model is low if there is a significant correlation between descriptors and randomized values of ΔG_{bind}. In other words, the

distance between the actual model (cross) and random models (dots) is an indicator of predictive power, i.e., lack of chance correlation. Note that the QMLIECE performs better in this test than LB, and much better than BD. This behavior is consistent with the decreasing physical soundness and increasing amount of fitting parameters in going from QMLIECE to LB and BD. The horizontal line at q^2=0.4 and the vertical line at R=0.6 are drawn to better compare the plots.

4 Thermodynamic cycle used in (QM)LIECE to calculate the binding free energy in solution



$$\Delta G_{elesol} = \Delta G^{vacuo}_{prot/lig} + \Delta G^{solvation}_{prot/lig} - \Delta G^{solvation}_{prot} - \Delta G^{solvation}_{lig}$$

Chapter 3

High-throughput Virtual Screening using Quantum Mechanical Probes: Discovery of Selective Kinase Inhibitors

Zhou, T.; Caflisch A. *ChemMedChem*, **2010**, in press

High-throughput Virtual Screening using Quantum Mechanical Probes: Discovery of Selective Kinase Inhibitors

Ting Zhou and Amedeo Caflisch*

May 25, 2010

Department of Biochemistry, University of Zurich,

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Phone: (+41 44) 635 55 21, Fax: (+41 44) 635 68 62

Email: caflisch@bioc.uzh.ch

*To whom correspondence should be addressed.

Abstract: A procedure based on the semi-empirical quantum mechanical (QM) calculation of the interaction energy is proposed for the fast screening of compound poses generated by high-throughput docking. Small molecules (consisting of 2–10 atoms and termed "probes") are overlapped to polar groups in the binding site of the protein target. The interaction energy values between each compound pose and the probes, calculated by a semi-empirical Hamiltonian, are used as filters. The QM probe method does not require fixed partial charges, and takes into account polarization and charge-transfer effects which are not captured by conventional force fields. The procedure is applied to screen about 100 million poses (of 2.7 millions of commercially available compounds) obtained by high-throughput docking in the ATP-binding site of the tyrosine kinase erythropoietin producing human hepatocellular carcinoma receptor B4 (EphB4). Three QM probes on the hinge region and one at the entrance pocket are employed to select for binding affinity, while a QM probe on the side chain of the so-called gatekeeper residue (a hypervariable residue in the kinome) is used to enforce selectivity. The poses with favorable interactions with the five QM probes are filtered further for hydrophobic matching and low ligand strain. In this way, a single-digit μM inhibitor of EphB4 with a relatively good selectivity profile is identified in a multi-million compound library upon experimental tests of only 23 molecules.

1 Introduction

Fast and accurate methods for computing the binding free energy between small molecules and proteins are required for computer-aided drug design.[1–6] The increasing popularity of QM methods in computer-aided drug design (CADD) is not just a consequence of ever growing computing power but is also due to the first principle nature of QM, which should provide the highest accuracy.[7–10] However, the computational time of QM ranges from N^3 (semi-empirical) to N^5 (second order Møller-Plesset perturbation theory and other post-Hartree-Fock methods), where N is the number of basis functions.[11] Therefore, QM is used for molecular systems of

limited size, e.g., in the hybrid quantum mechanics/molecular mechanics approach[12,13], or for a small subset of atoms while a polarizable continuum model is employed to describe the protein and the solvent.[14] In linear scaling QM method, the computing time scales with N^2 or even N if the local character of chemical interactions is fully exploited.[7,15–19] Using linear scaling theory, Stewart and Anikin et al. applied the software package MOZYME[20] and LocalSCF[21], respectively, to calculate QM energies with localized molecular orbital (LMO) theory. Recently, we have developed the quantum mechanical linear interaction energy model with continuum electrostatic solvation (QMLIECE), in which the electrostatic contribution to the binding energy is evaluated by a semi-empirical QM divide-and-conquer strategy. QMLIECE is useful for highly variable charge–charge interactions, as in the case of 44 peptidic inhibitors of a flaviviral non-structural 3 serine protease.[22] Nevertheless, neither LMO theory nor the QMLIECE approach are fast enough for evaluating multiple poses of small molecules generated by high-throughput virtual screening (HTVS). Therefore, Vasilyev and Bliznyuk first used fast scoring functions for ranking and then applied the LMO theory, as implemented in MOZYME, to only the 10–100 top ranking poses.[23] We have applied QMLIECE to less than 10000 poses at most.[22] No application of QM methods to millions of poses in HTVS has been reported as of today.

Here, a procedure based on semi-empirical QM is developed for the in silico screening of millions of poses generated by high-throughput docking of large libraries of compounds. The interaction energies (IEs) between small molecules and individual polar groups (probes) in the binding pocket of the protein target are used to filter out poses that are not likely to bind. The method focuses on exploiting advantages of QM in HTVS, e.g., independence of force field parameters, and ability to capture charge transfer, polarization, and direction-specific effects which are very important for describing hydrogen bonds (HBs).[24] Flexibility of the functional groups of some of the side chains is taken into account by partial optimization of the structure of the complex. As a proof of principle, the QM probe approach is applied to the receptor tyrosine kinase EphB4, which is involved in cancer-related angiogenesis.[25–27] Docking is performed at the

ATP-binding site, and five QM probes are used for filtering: Three probes represent the backbone polar groups of the hinge region[28], one probe is located at the entrance pocket, and a fifth probe is selected at the gatekeeper side chain[28–32] (Thr693 in EphB4) to bias the in silico screening towards selective inhibitors of EphB4.

2 Methods

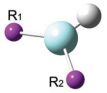
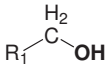
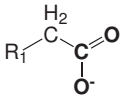
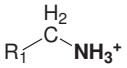
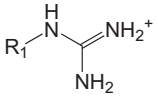
2.1 Design of QM probes

The QM probes are small molecular fragments (2–10 atoms) used for the efficient evaluation of the IE between polar groups of the protein and the ligand. The QM probes are "designed" (1) to reflect the local electronic structures in the binding pocket, (2) to be as simple as possible for computational efficiency, and (3) to distinguish between favorable and adverse contact sensitively. Methanol, acetate anion, methylammonium cation, and guanidinium cation probes are essentially identical to the corresponding functional groups in the side chains of Ser/Thr/Tyr, Asp/Glu, Lys, and Arg, respectively (Table 1).

A water molecule is used as probe for the carbonyl group in the backbone as well as in the Asn and Gln side chains. Furthermore, the plane of the water molecule is perpendicular to the carbonyl group to have the same orientation of the lone pairs (Table 1).[33,34] Water is more appropriate than acetone or acetamide because the additional methyl or amino group, respectively, may be involved in van der Waals (vdW) interactions with the ligand, and thus attenuate the energetic difference due to the formation of HB. Formaldehyde is not as sensitive as the water molecule to detect the formation of HB, because it has a carbon atom more than water, and the carbon atom may also form vdW interaction with the ligand.

The hydrogen fluoride (HF) probe is used to emulate the amide group in the backbone of the protein as well as in the Asn and Gln side chains for two reasons. First, HF is the strongest neutral HB donor, and thus is the most sensitive as a probe molecule to detect HB acceptors. Second, HF

Table 1: The QM probes for polar groups in proteins. The atoms in boldface are flexible during QM minimization of complex formation enthalpy, which is carried out with rigid ligand.

Functional group	Location in proteins	Probe molecule	Overlapping position
	backbone, Asn, Gln	H_2O	
	backbone, Asn, Gln	HF	
	Ser, Thr, Tyr	$\mathbf{CH_3OH}$	—
	Asp, Glu	$\mathbf{CH_3CO^{2-}}$	—
	Lys	$\mathbf{CH_3NH_3^+}$	—
	Arg	$C(NH_2)_3^+$	—

is a two-atom molecule, and has little additional vdW interaction with ligands.

The QM probe method was first assessed on cyclin-dependent kinase 2 (CDK2) (see Supp. Mat.) and then applied to EphB4. The application to CDK2 shows that the method is able to identify classical HBs as well as favorable polar interactions like the one between aromatic CH and carbonyl oxygen[35].

2.2 Calculation of interaction energy

Multiple poses of the ligands are determined by automatic docking and force field minimization (see below). Ligands are always fixed during the evaluation of the interaction energy IE $=$ $H_{\text{ligand-probe}} - H_{\text{ligand}} - H_{\text{probe}}$, where H is the formation enthalpy calculated with MOPAC[36] and the semi-empirical Hamiltonian PM6.[37] It has been reported that density functionals can quantitatively reproduce the HB energy of CCSD(T) (Coupled-Cluster with Single and Double and perturbative Triple excitations),[38–44] but considering the efficiency required for filtering multi-million poses, rapid PM6 was selected eventually.[45] The IEs between rigid probes and the ligand are calculated directly, while optimization of $H_{\text{ligand-probe}}$ is carried out for flexible probes. In particular, the coordinates of the atoms in boldface in Table 1 are optimized to find an energy minimum of the ligand-probe complex. The IE evaluation of rigid probes is fast (less than half second), while about 30 seconds are required for the flexible probes. To improve efficiency, the interactions of the ligand with the rigid probes are calculated first and can be used as filters. This strategy was used in the application to EphB4 (see subsection 3.2).

2.3 Preparation of EphB4

Since the structure of the kinase domain of EphB4 was not available when we started this work, a homology model was built using the structure of EphB2 (mouse, PDB entry 1JPA) as template. The sequence identity between human EphB4 and mouse EphB2 is 88%. A detailed description

of the homology modelling procedure has been published previously[46]. Notably, the root mean square deviation (RMSD) between the homology model and the crystal structure of EphB4 (PDB entry 2VWX) is only 0.28 Å for 179 of 242 C_α atoms. Moreover, the orientation of the side chains in the ATP-binding site is essentially identical in the model and the X-ray structure, the largest discrepancy (1.1 Å) being at the tip of the Met668 side chain in the hydrophobic pocket.

2.4 Preparation of the library and initial filtering

Of the 9.8 million compounds in the 2007 version of the ZINC library[47], about 2.7 millions had at least one HB donor and one HB acceptor, and molecular weight (MW) smaller than 500 $g \cdot mol^{-1}$. The molecules were assigned protonation states at pH=7, and were prepared in multiple protonation states and multiple tautomeric forms. The presence of HB donor(s) and acceptor(s) is essential as the QM probe method focuses on the evaluation of HB patterns[35], while the filter on MW was employed because small molecules are more appropriate as lead compounds. The molecular properties used for filtering were calculated by DAIM.[48] The program WITNOTP (Armin Widmer, Novartis Pharma, Basel) was used for automatic assignment of CHARMm atom types and parameters[49], including partial charges which were determined by an iterative approach based on the partial equalization of orbital electronegativity.[50,51]

2.5 Docking into EphB4

Version 4 of AutoDock[52] was used for flexible ligand docking of the ZINC subset of 2.7 million compounds using a rigid protein. First, the atom-specific affinity map files were generated by AutoGrid.[53] The numbers of points in the x, y, and z directions were 62, 52, and 42, respectively, and the spacing between two adjacent grid points was 0.25 Å. Then AutoDock was employed to generate multiple poses for further minimization by CHARMM[54]. Since the AutoDock scoring function was not used for ranking, to speed up the docking procedure, AutoDock energy

evaluations was limited to 25000 for each hybrid-genetic-local-search. The hybrid-genetic-local-search was run 400 times with different initial seeds to obtain multiple poses for each compound. (In preliminary docking runs of known inhibitors of EphB4, it was more likely to obtain the correct binding mode by using a large number of hybrid-genetic-local-searches than a large number of energy evaluations and only a few searches.) The docking was followed by energy minimization in the rigid protein using the CHARMm force field.[49] Redundant poses were eliminated by clustering using an all-atom RMSD cutoff 0.01 Å. For each pose, the IE values with the QM probes as well as electrostatics and vdW efficiencies were stored in a table of DVSDMS (data management system for distributed virtual screening).[55]

2.6 van der Waals filters

Upon energy minimization, loose vdW filters[56] were applied to all poses to eliminate those with clashes and/or poor steric complementarity.[7] The following cutoffs of the CHARMm intermolecular vdW energy were employed: $E_{\text{vdW}} < -20$ kcal $\cdot$ mol^{-1} and $E_{\text{vdW}}/\text{MW} < -0.05$ kcal $\cdot$ g^{-1}. Considering the efficiency and accuracy, the vdW filters are more appropriately calculated with force field methods than with QM.[57]

3 Results and Discussion

3.1 Docking and van der Waals filters

A total of about 100 million poses of 2.7 million compounds were generated by flexible ligand docking into the rigid ATP-binding site of EphB4. The vdW filters reduced the number of poses to about 90 millions.

3.2 Polar interactions: QM probe energies

Thirteen polar groups were replaced by QM probes in the ATP-binding site of EphB4 (Figure 1). According to the definitions of Traxler and Furet,[28,58] Probes 1–3 lie in the adenine binding region, Probe 4 is located at the entrance pocket, Probes 5 and 9 lie in the ribose binding pocket, Probe 6 is in the phosphate binding pocket, and Probes 10–12 lie in the hydrophobic pocket. The virtual screening focuses on binding at the hinge region, therefore Probes 1–3 were selected for filtering. Probe 4 was also taken into account for filtering as it is very close to the hinge region. Moreover, the hydroxyl group of the gatekeeper residue (Probe 10 at Thr693) was included to bias the search towards selective kinase inhibitors[28–32], as only 95 of the 507 human protein kinases have Thr as a gatekeeper.[59] Whether to consider a specific probe depends on the requirement, e.g., Probe 5 and 9 have to be taken into account if the polar interactions in the ribose binding pocket need to be inspected. Therefore the unused probes may be employed in future studies, e.g., for combinatorial lead optimization or de novo design.[60]

Four known inhibitors of EphB4 were used to determine the cutoffs for filtering according to the QM probe energy. The four inhibitors were docked into the ATP-binding site of EphB4 and minimized with the same protocols as used for the high-throughput docking. Table 2 shows structures and probe energies of these inhibitors. The probe energy values are robust upon minor shifts in the binding mode (see Suppl. Mat. Table S-I).

The QM probe energies of the four known EphB4 inhibitors and about 100 compounds (selected randomly as representative of inactive compounds) were taken into account to determine the cutoffs for filtering. These are: $E_{\text{probe1}} < -2$ kcal $\cdot$ mol^{-1} (typical hydrogen bond energy of N$-$H$\cdots$: O)[64] and $E_{\text{probe2}}, E_{\text{probe3}}, E_{\text{probe4}}$, and $E_{\text{probe10}} < 0.5$ kcal $\cdot$ mol^{-1}. Furthermore, the filter based on the sum $E_{\text{probe1}} + E_{\text{probe2}} < -3.8$ kcal $\cdot$ mol^{-1} was employed to give more weight to the two most buried polar groups of the hinge region. Since the methanol probe 10 at the gatekeeper residue is flexible, its interaction was calculated only if the $E_{\text{probe1}} < 0$ kcal $\cdot$ mol^{-1}. The filters based on QM probes reduced the number of poses from 90 millions to 955,094 poses

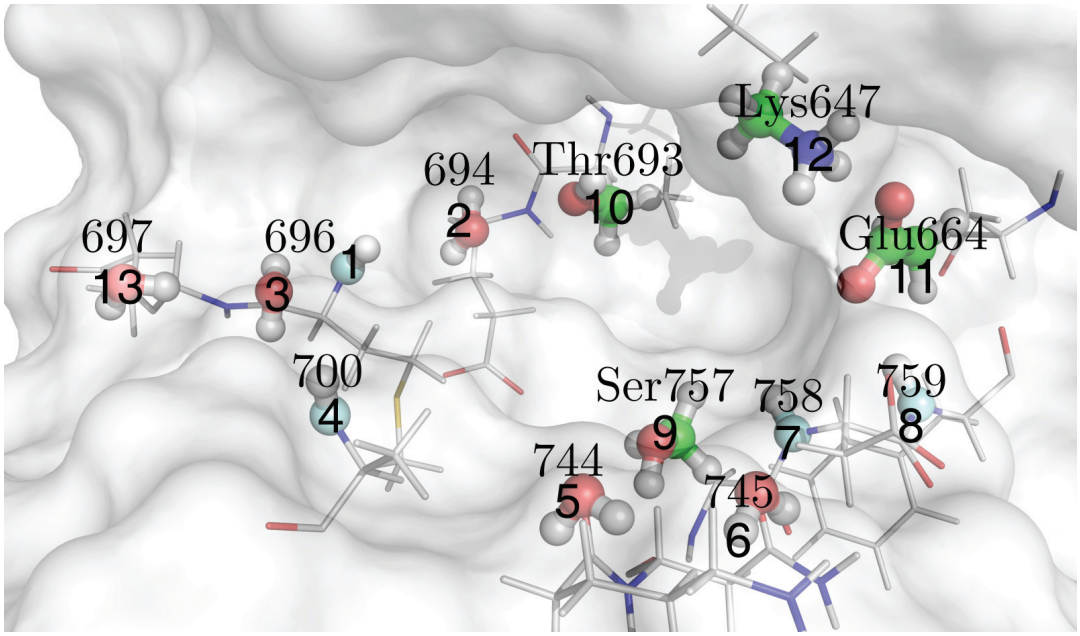
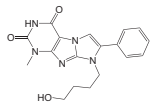
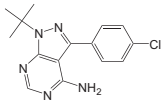
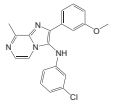
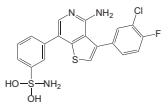


Figure 1: The 13 QM probes in the ATP-binding site of the receptor tyrosine kinase EphB4. The probes are shown by spheres colored according to the atomic element with hydrogen in gray, carbon in green, nitrogen in blue, oxygen in red, and fluorine in cyan. The QM probes on side chains are labeled with the residue type and number, while those on the backbone are labeled only by the residue number.

Table 2: Values of QM probe energies and experimentally measured IC_{50} of four known inhibitors of EphB4. The probe energies are in kcal·mol^{-1}.

Compound	Structure	MW (g·mol^{-1})	IC_{50} (μM)	Probe 1	Probe 2	Probe 3	Probe 4	Probe 10
ALTA2[46]		353	1.4	−4.31	0.37	0.14	0.34	0.41
PP2[61]		302	0.34	−3.71	−3.18	−1.24	0.12	−2.98
ONC102[62]		365	0.10	−2.46	−2.43	−1.38	−0.22	−2.32
MIYA9f[63]		434	0.021	−3.04	−2.25	−1.16	−5.53	−3.47

of 509,101 molecules. In other words, the QM probe filtering diminished the average number of poses per molecule from 37 to 2.

3.3 Apolar interactions: Hydrophobic matching

The QM probes do not account for hydrophobic surface matching upon binding, which is important to evaluate the interaction at the hydrophobic pocket for kinases with small gatekeeper residue (e.g., EphB4).[65] The atoms of the ligand were classified as polar or nonpolar according to their QM-calculated partial charges (semi-empirical PM6 Hamiltonian and Mulliken population analysis). By comparing the QM charges of typical polar and nonpolar atoms in small molecules, 0.22 electronic units was selected as threshold, i.e., those atoms with partial charge in the range from -0.22 to 0.22 electronic units were considered apolar, while the remaining atoms polar. Although this assignment requires the choice of an arbitrary threshold value, it was adopted because of its efficiency and simplicity. The hydrophobic matching was approximated by the

vdW interactions between the residues within the hydrophobic pocket of EphB4 (Val629, Ala645, nonpolar part of Lys647, Met668, Ile691, and Thr693) and the nonpolar atoms of the ligand. A hydrophobic matching of at least -5 kcal·mol^{-1}was used to significantly reduce the number of poses for further analysis (15,979 poses belonging to 13,823 molecules). Visual inspection of some of the discarded poses confirmed that they do not fill the hydrophobic pocket with significant apolar surface matching.

3.4 Ligand strain filter

To evaluate the strain of the ligand, a minimization was performed in the absence of the protein starting from the bound conformation. The program MOPAC with a semi-empirical Hamiltonian RM1[66] was used for the minimization. The program ROCS[67] was employed to overlap the conformation minimized in the absence of the protein to the pose used as the starting point of the minimization, and to calculate the shape Tanimoto. The latter is defined as $O_{\mathrm{AB}}/(V_{\mathrm{A}} + V_{\mathrm{B}} - O_{\mathrm{AB}})$, where O_{AB} is the volume overlap between conformer A and conformer B, and V_{A} and V_{B} is the volume of conformer A and B, respectively.[68] A shape Tanimoto close to 1 implies that the two conformations are essentially identical. The distribution of the shape Tanimoto of the 15,979 poses is shown in Suppl. Mat. Figure S-VII. A threshold for shape Tanimoto larger than 0.9 was chosen to further reduce the amount of poses for visual inspection. In total, 8,461 poses belonging to 7,536 molecules passed this filter. Finally, 23 compounds were selected for experimental validation upon visual inspection of the first 1000 poses sorted according to QM probe energies and CHARMm intermolecular energy (See Suppl. Mat. Table S-IV). Sortings were simply carried out according to the sum of CHARMm intermolecular non-bonding energy terms, and the five probe energies (Probe 1–4 and Probe 10) separately. The top ~ 166 poses of each ranking were selected. The main criteria used to filter out poses during visual inspection were the involvement of highly flexible functional groups, e.g., $-(\mathrm{CH_2})_n-\mathrm{OH}$ ($n \geq 1$), in HBs with the hinge region, and the desolvation of polar groups in the ATP-binding site that were not involved in intermolecular

HBs. The former was required as the QM probe filters do not take into account the conformational entropy loss upon binding. Visual inspection, although subjective, is an unavoidable procedure for discarding poses with unfavorable interactions and/or unlikely conformations[69]. Filtered by the QMprobe and vdW energies, the remaining poses are inspected very efficiently, as key polar and apolar interactions have already been verified.

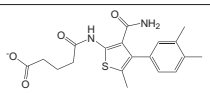
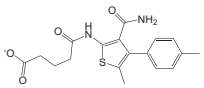
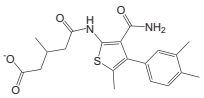
3.5 Computational requirements

The docking approach requires about 10 minutes per compound on a single Opteron CPU 244 (2.4GHz). The energy minimization requires about 10 minutes for an average of 50 poses for each compound. The CPU time for calculating an IE with a rigid QM probe is less than 0.5 second mainly for the program input/output process, while about 30 seconds are needed for calculating the IE with a flexible probe. Despite the large amount of molecules and poses, distributing docking and minimization jobs to hundreds of CPUs in two Beowulf clusters, selecting specific molecules and poses, and applying filters was efficiently managed by the database management system DVSDMS.[55]

3.6 Experimental validation

The 23 compounds selected for validation were tested in two different enzymatic assays with the kinase catalytic domain of EphB4 in solution. One assay is based on fluorescence-resonance energy transfer (FRET) between coumarin and fluorescein (Omnia® Tyr recombinant kit KNZ4051, BioSource™[70]), while the other measures the amount of radioisotope labeled phosphate transferred from ATP to the substrate upon phosphorylation by EphB4.[71] Three of the 23 compounds tested share a 2-formamido-4-phenylthiophene-3-carboxamide scaffold. Two of these three thiophene derivatives have an inhibitor concentration for half-maximal activity (IC_{50}) smaller than 10 μM in both assays, while the third one has IC_{50} values of 9 μM and 17 μM (Table 3).

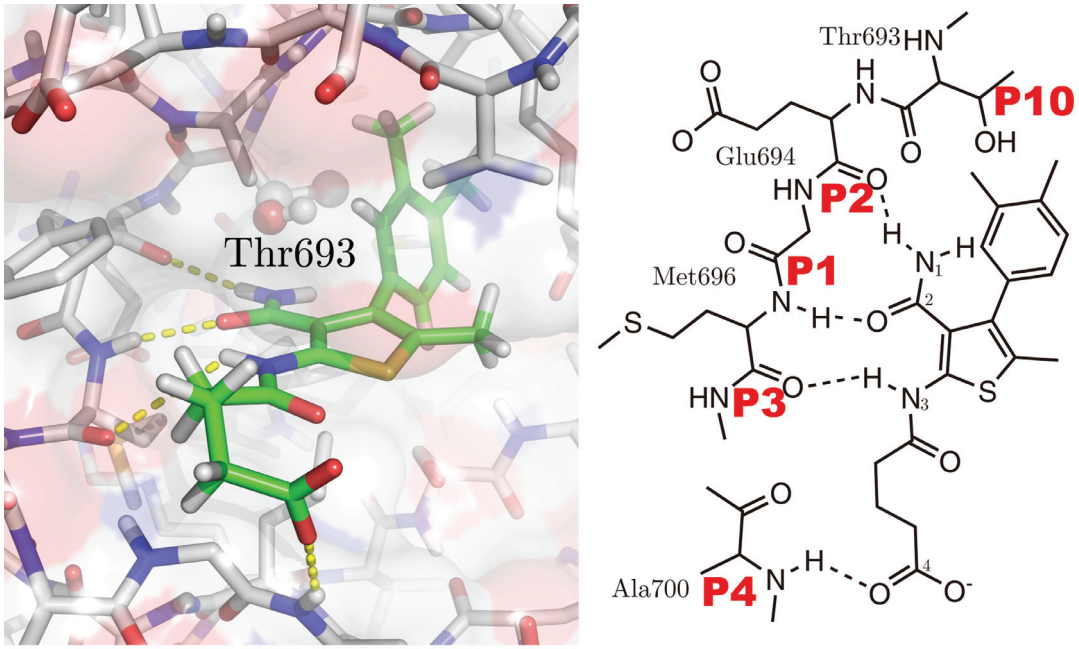
Table 3: Experimental validation of EphB4 inhibitors identified by high-throughput docking and the QM probe approach.

Compound	Structure	MW ($g \cdot mol^{-1}$)	IC_{50} [a] (μM)	IC_{50} [b] (μM)
1		373	8.4, 7.6	2
2		360	7.1, 4.6	2
3		388	17.5, 16.7	9

[a] FRET-based enzymatic assay with the recombinant catalytic domain of human EphB4 in solution and ATP concentration of 20 μM. Each inhibitor was tested twice. To provide evidence against non-specific effects (e.g., aggregation), compound **2** was also tested upon addition of the detergent triton X-100 (0.1% v/v). Similar values of percentage inhibition (at 30 μM and 10 μM of compound **2**) were measured with and without detergent. [b] Enzymatic assay with the recombinant catalytic domain of human EphB4 and [γ-^{33}P]-ATP concentration of 1 μM (performed at Reaction Biology Corp).

The predicted binding mode of compound **1** (Figure 2) indicates that its amide group N_1-H, carbonyl group $C_2=O$, and amide group N_3-H are involved in HBs with the backbone polar groups in the hinge region. Moreover, its carboxy group is involved in a HB with the entrance pocket, and the dimethylphenyl ring is buried into the hydrophobic pocket. To validate the predicted binding mode of compound **1**, a set of 23 commercially available derivatives with the same 2-formamido-4-phenylthiophene-3-carboxamide scaffold were purchased and tested (Table 4). The SAR (structure-activity relationship) data are consistent with the binding mode. In particular, the cyclopropyl group at R_7 causes a major loss in activity (compound **12**) in agreement with the hinge region hydrogen bond of the amide group N_1-H. Moreover, the similar inhibitory activity of compounds **22** and **23** (about 50% @20 μM concentration), which differ only by a methyl group at R_2, is consistent with the orientation of the R_2 substituent towards the solvent (Figure 2).

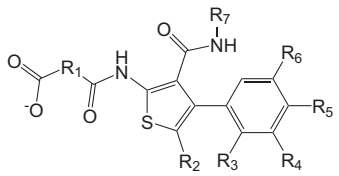
The selectivity profile of compound **1** was tested using a panel of 85 protein kinases (National Centre for Protein Kinase Profiling at University of Dundee, see Suppl. Mat. Table S-II). At 10 μM concentration, the activity of Aurora B remained 37% compared with a DMSO control, while six other kinases retained $40-60\%$ activity (Table 5). Note that three of these seven kinases have Thr as gatekeeper residue. Importantly, the inhibitory activity of compound **1** on the remaining 78 kinases is either zero or extremely modest. The binding mode obtained by docking into EphB4 suggests that the phenyl ring of compound **1** interacts favorably with the hydroxyl group of the gatekeeper Thr693 ($E_{\text{probe10}} = -1.6$ kcal·mol^{-1}). These results indicate that compound **1** is rather selective which is in part due to the use of the QM probe 10 at the gatekeeper side chain (only about 20% of human kinases have Thr as gatekeeper).[59] Interestingly, the known inhibitors PP2[61], ONC102,[62] and MIYA9f[63] have very favorable interaction energy with QM probe 10 (Table 2) which is consistent with their good selectivity for protein kinases with Thr as gatekeeper.



(a) (b)

Figure 2: Binding mode of compound **1** predicted by docking. (a) The intermolecular HBs to the hinge loop and entrance loop are shown by yellow dashed lines. The pose was minimized in the rigid EphB4 structure (PDB entry 2VWX). The atoms of the side chain of the gatekeeper residue Thr693 are displayed by spheres. Non-polar hydrogen atoms are omitted for clarity. (b) The five QM probes used as filters are emphasized by red labels. Integer labels on compund **1** emphasize functional groups mentioned in the text. The side chain of Phe695 is not shown for clarity.

Table 4: Structure and inhibitory activity of 23 commercially available derivatives of compound **1**.



Compound	R_1	R_2	R_3	R_4	R_5	R_6	R_7	$IC_{50}(\mu M)$ or % inhibitory activity @20μM [a]
4	$-CH_2-CH_2-$	$-CH_3$	H	H	F	H	H	49%@20μM
5		$-CH_3$	H	H	H	H	H	50%@20μM
6	$-CH_2-CH_2-CH_2-$	$-CH_3$	H	H	$-C(CH_3)_3$	H	H	38%@20μM
7	$-CH_2-CH_2-$	H	H	$-CH_3$	$-CH_3$	H	H	46%@20μM
8		H	$-CH_3$	H	H	$-CH_3$	H	50%@20μM
9		H	Cl	H	Cl	H	H	9.7
10		$-CH_3$	H	H	$-CH_3$	H	H	13.7
11		$-CH_3$	Cl	H	Cl	H	H	7.8
12	$-CH_2-CH_2-CH_2-$	H	H	H	$-CH_3$	H		7%@20μM
13		$-CH_3$	H	H	$-CH_3$	H	H	40%@20μM
14		H	H	$-CH_3$	$-CH_3$	H	H	44%@20μM
15		$-CH_3$	H	$-CH_3$	$-CH_3$	H	H	31%@20μM
16		H	H	$-CH_3$	$-CH_3$	H	H	21%@20μM
17		$-CH_3$	H	$-CH_3$	$-CH_3$	H	H	16%@20μM
18	$-CH_2-CH_2-CH_2-$	H	H	$-CH_3$	$-CH_3$	H	H	55%@20μM
19	$-CH_2-CH_2-CH_2-$	$-CH_3$	H	H	$-CH_2-CH_3$	H	H	16.3
20		$-CH_3$	H	$-CH_3$	$-CH_3$	H	H	27%@20μM
21		H	H	H	$-CH_3$	H	H	15%@20μM
22	$-CH_2-CH_2-CH_2-$	$-CH_3$	Cl	H	Cl	H	H	47%@20μM
23	$-CH_2-CH_2-CH_2-$	H	Cl	H	Cl	H	H	54%@20μM
24		$-CH_3$	H	H	$-CH_3$	H	H	5.2
25		$-CH_3$	H	$-CH_3$	$-CH_3$	H	H	2%@20μM
26		H	H	H	Cl	H	H	20%@20μM

[a] FRET-based enzymatic assay with the recombinant catalytic domain of human EphB4 in solution and ATP concentration of 20 μM. Each inhibitor was tested twice and average values are reported.

Table 5: Selectivity profile of compound **1** tested on a panel of 85 protein kinases.

Kinase	% activity remaining at 10 μM conc. of compound **1** [a]	Gatekeeper residue [b]
Aurora B	37	Leu
EPH A2	41	Thr
VEG-FR	42	Thr
P38B MAPK	51	Thr
FGF-R1	60	Val
LKB1	60	Met
CK1δ	60	Met
5 kinases	$60 < \%\text{activity} \leq 70$	
13 kinases	$70 < \%\text{activity} \leq 80$	
10 kinases	$80 < \%\text{activity} \leq 90$	
Other 50 kinases	> 90	

The measurements of inhibitory activity were performed at the National Centre for Protein Kinase Profiling at University of Dundee. [a] The activity is relative to a control with 100% of DMSO. The 78 unnominated kinases have the following residue as gatekeeper with occurrence in parentheses: Met (35), Leu (16), Phe (12), Thr (10), Gln (2), Glu (1), Ile (1), and Tyr (1). The sequence information and gatekeeper residue of the whole panel of 85 kinases is shown in the Suppl. Mat. (Table S-III). The ATP concentration in the assay is also shown in the Suppl. Mat. (Table S-II).

4 Conclusions

A new procedure for filtering millions of poses of small molecules based on QM calculations is described by an application to the tyrosine kinase EphB4. Polar groups in the protein binding site are substituted by small probes consisting of 2–10 atoms, e.g., water and methanol for the carbonyl and hydroxyl group, respectively, and the interaction energy values between each pose and individual QM probes are calculated using a semi-empirical Hamiltonian. The use of a first-principle method is an advantage with respect to classical force fields with fixed partial charges. As an example, the QM probes are able to detect favorable polar interactions like the aromatic $-$CH$\cdots$O=C interaction, which can be rather strong depending on the electronegativity of eventual substituents in the aromatic ring.

The QM probe filtering is applied to about 100 millions poses of small molecules generated by automatic docking into the ATP-binding site of the kinase catalytic domain of EphB4. Only 1% of the poses pass the filters of favorable interaction energy with five QM probes: three in the hinge region, one at the entrance pocket, and one at the gatekeeper side chain (Thr693 in EphB4). The latter is a non-conserved residue in the kinome, and is therefore used to bias the virtual screening towards selective inhibitors. Upon further filtering based on nonpolar interactions, ligand strain, and visual inspection, 23 compounds are selected and tested in an enzymatic assay. It is important to note that the QM probe filters as well as the additional filters used for post-processing require rather arbitrary threshold values, which might seem in contradiction with the use of first-principle methods. The QM probe method is an approximation of the real binding free energy because it takes into account only part of the protein target. Moreover, entropic terms are neglected.

Of the 23 compounds tested, three molecules with a 2-formamido-4-phenylthiophene-3-carboxamide scaffold are active in the low μM range in two different enzymatic assays. Additional evidence for the binding mode of compound **1**, and in particular its favorable interactions with the protein functional groups approximated by the QM probes, is provided by the structure-activity

relationship of 25 commercially available derivatives. Enzymatic assays on a panel of 85 protein kinases indicate that compound **1** is not promiscuous as no inhibitory activity is observed for most of these kinases and modest inhibitory activity for only seven of them (three of which have Thr as gatekeeper residue). Thus, compound **1** has potential for further development into a lead candidate, because of its low μM inhibitory activity for EphB4, low MW (373 g·mol^{-1}), and good selectivity profile.

Acknowledgement. We thank Dr. Danzhi Huang for performing some of the enzymatic assays as well as for interesting discussions and comments to the manuscript. We thank Dr. Philipp Schütz and Christian Bolliger for maintaining the Beowulf Linux cluster Etna and Matterhorn, respectively, which were used for most of the computations. We are grateful to Armin Widmer for providing the modelling program WITNOTP which was used for generating topology files and assigning CHARMm atom types. This work was supported by grants of the Swiss National Science Foundation to A.C.

Support Information. The supporting information includes assessment of QM probe method on the protein kinase CDK2, probe energies of minor-different conformers of four known inhibitors of EphB4, the distributions of energies of 5 probes (Probe 1–4, and 10), the IE against distances between a water molecule and a N-methylacetamide calculated by PM6 Hamiltonian, the distribution of shape Tanimoto of 15,979 poses, and a list of kinases for screening assay.

References

[1] J. Apostolakis, A. Caflisch, *Comb. Chem. High Throughput Screening* **1999**, *2*, 91–104.

[2] R. C. Glen, S. C. Allen, *Curr. Med. Chem.* **2003**, *10*, 763–777.

[3] W. P. Walters, M. Namchuk, *Nat. Rev. Drug Discov.* **2003**, *2*, 259–266.

[4] D. Huang, A. Caflisch, *J. Med. Chem.* **2004**, *47*, 5791–5797.

[5] W. Jorgensen, *Science* **2004**, *303*, 1813–1818.

[6] M. Wang, C. F. Wong, *J. Chem. Phys.* **2007**, *126*, 026101.

[7] K. Raha, K. M. Merz, *J. Med. Chem.* **2005**, *48*, 4558–4575.

[8] A. Cavalli, P. Carloni, M. Recanatini, *Chem. Rev.* **2006**, *106*, 3497–3519.

[9] M. B. Peters, K. Raha, K. M. Merz, *Curr. Opin. Drug Discov. Devel.* **2006**, *9*, 370–379.

[10] T. Zhou, D. Huang, A. Caflisch, *Curr. Top. Med. Chem.* **2010**, *10*, 33–45.

[11] A. Van der Vaart, V. Gogonea, S. Dixon, K. Merz, *J. Comput. Chem.* **2000**, *21*, 1494–1504.

[12] H. M. Senn, W. Thiel, *Angew. Chem. Int. Ed. Engl.* **2009**, *48*, 1198–1229.

[13] P. Fong, J. P. McNamara, I. H. Hillier, R. A. Bryce, *J. Chem. Inf. Model.* **2009**, *49*, 913–924.

[14] V. Buback, M. Mladenovic, B. Engels, T. Schirmeister, *J. Phys. Chem. B* **2009**, *113*, 5282–5289.

[15] S. Gadre, R. Shirsat, A. Limaye, *J. Phys. Chem.* **1994**, *98*, 9165–9169.

[16] S. Dixon, K. Merz, *J. Chem. Phys.* **1996**, *104*, 6643–6649.

[17] T. Lee, J. Lewis, W. Yang, *Comp. Mat. Sci.* **1998**, *12*, 259–277.

[18] Y. R. Mo, J. L. Gao, S. D. Peyerimhoff, *J. Chem. Phys.* **2000**, *112*, 5530–5538.

[19] D. W. Zhang, Y. Xiang, A. M. Gao, J. Z. Zhang, *J. Chem. Phys.* **2004**, *120*, 1145–1148.

[20] J. J. P. Stewart, *Int. J. Quantum Chem.* **1996**, *58*, 133–146.

[21] N. A. Anikin, V. M. Anisimov, V. L. Bugaenko, V. V. Bobrikov, A. M. Andreyev, *J. Chem. Phys.* **2004**, *121*, 1266–1270.

[22] T. Zhou, D. Huang, A. Caflisch, *J. Med. Chem.* **2008**, *51*, 4280–4288.

[23] V. Vasilyev, A. Bliznyuk, *Theor. Chem. Acc.* **2004**, *112*, 23.

[24] R. A. Friesner, *Adv. Protein Chem.* **2005**, *72*, 79–104.

[25] R. H. Adams, *Semin. Cell Dev. Biol.* **2002**, *13*, 55–60.

[26] G. Martiny-Baron, T. Korff, F. Schaffner, N. Esser, S. Eggstein, D. Marmé, H. G. Augustin, *Neoplasia* **2004**, *6*, 248–257.

[27] N. Kertesz, V. Krasnoperov, R. Reddy, L. Leshanski, S. R. Kumar, S. Zozulya, P. S. Gill, *Blood* **2006**, *107*, 2330–2338.

[28] P. Traxler, P. Furet, *Pharmacol. Ther.* **1999**, *82*, 195–206.

[29] M. E. M. Noble, J. A. Endicott, L. N. Johnson, *Science* **2004**, *303*, 1800–1805.

[30] P. J. Alaimo, Z. A. Knight, K. M. Shokat, *Bioorg. Med. Chem.* **2005**, *13*, 2825–2836.

[31] J. J. Liao, *J. Med. Chem.* **2007**, *50*, 409–424.

[32] M. Azam, M. A. Seeliger, N. S. Gray, J. Kuriyan, G. Q. Daley, *Nat. Struct. Mol. Biol.* **2008**, *15*, 1109–1118.

[33] J. A. Odutola, T. R. Dyke, *J. Chem. Phys.* **1980**, *72*, 5062–5070.

[34] P. Bobadova-Parvanova, B. Galabov, *J. Phys. Chem. A* **1998**, *102*, 1815–1819.

[35] S. Sarkhel, G. R. Desiraju, *Proteins* **2004**, *54*, 247–259.

[36] J. J. P. Stewart, *J. Comput. Chem.* **1989**, *10*, 209–220.

[37] J. J. P. Stewart, *J. Mol. Model.* **2007**, *13*, 1173–1213.

[38] H. G. Korth, M. I. de Heer, P. Mulder, *J. Phys. Chem. A* **2002**, *106*, 8779–8789.

[39] N. Turki, A. Milet, O. Ouamerali, R. Moszynski, E. Kochanski, *THEOCHEM.* **2002**, *577*, 239–253.

[40] R. A. Klein, *J. Comput. Chem.* **2003**, *24*, 1120–1131.

[41] J. Ireta, J. Neugebauer, M. Scheffler, *J. Phys. Chem. A* **2004**, *108*, 5692–5698.

[42] Y. Zhao, O. Tishchenko, D. G. Truhlar, *J. Phys. Chem. B* **2005**, *109*, 19046–19051.

[43] Y. X. Wang, B. Paulus, *Chem. Phys. Lett.* **2007**, *441*, 187–193.

[44] H. R. Leverentz, D. G. Truhlar, *J. Phys. Chem. A* **2008**, *112*, 6009–6016.

[45] *Comparison of PM6 and X-Ray Structures of Hydrogen-Bonded systems*, accessed on Nov. 17, 2009. {http://openmopac.net/Hydrogen\%20bonds.html}.

[46] P. Kolb, C. B. Kipouros, D. Huang, A. Caflisch, *Proteins* **2008**, *73*, 11–18.

[47] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

[48] P. Kolb, A. Caflisch, *J. Med. Chem.* **2006**, *49*, 7384–7392.

[49] F. Momany, R. Rone, *J. Comput. Chem.* **1992**, *13*, 888–900.

[50] K. No, J. Grant, H. Scheraga, *J. Phys. Chem.* **1990**, *94*, 4732–4739.

[51] K. No, J. Grant, M. Jhon, H. Scheraga, *J. Phys. Chem.* **1990**, *94*, 4740–4746.

[52] D. S. Goodsell, A. J. Olson, *Proteins* **1990**, *8*, 195–202.

[53] *AutoGrid*, accessed on July, 2, 2009. http://autodock.scripps.edu/wiki/AutoGrid.

[54] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig,

S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comput. Chem.* **2009**, *30*, 1545–1614.

[55] T. Zhou, A. Caflisch, *J. Chem. Inf. Model.* **2009**, *49*, 145–152.

[56] P. Kolb, D. Huang, F. Dey, A. Caflisch, *J. Med. Chem.* **2008**, *51*, 1179–1188.

[57] T. Giese, D. York, *Int. J. Quantum Chem.* **2004**, *98*, 388–408.

[58] M. Cherry, D. H. Williams, *Curr. Med. Chem.* **2004**, *11*, 663–673.

[59] D. Huang, T. Zhou, K. Lafleur, C. Nevado, A. Caflisch, *Bioinformatics* **2010**, *26*, 198–204.

[60] F. Dey, A. Caflisch, *J. Chem. Inf. Model.* **2008**, *48*, 679–690.

[61] A. Sturz, B. Bader, K. H. Thierauch, J. Glienke, *Biochem. Biophys. Res. Commun.* **2004**, *313*, 80–88.

[62] C. Berset, S. V. Audetat, A. Barberis, T. Gunde, J. Tietz, P. Traxler, A. Schumacher, *Protein kinase inhibitor*, **U.S. Patent 2007/0149535**. `http://www.google.com/patents/about?id=wqigAAAAEBAJ`.

[63] Y. Miyazaki, M. Nakano, H. Sato, A. T. Truesdale, J. D. Stuart, E. N. Nartey, K. E. Hightower, L. Kane-Carson, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 250–254.

[64] H. Adalsteinsson, A. H. Maulitz, T. C. Bruice, *J. Am. Chem. Soc.* **1996**, *118*, 7689–7693.

[65] N. G. Ahn, K. A. Resing, *Science* **2005**, *308*, 1266–1267.

[66] G. B. Rocha, R. O. Freire, A. M. Simas, J. J. P. Stewart, *J. Comput. Chem.* **2006**, *27*, 1101–1111.

[67] *ROCS*, OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

[68] J. Grant, M. Gallardo, B. Pickup, *J. Comput. Chem.* **1996**, *17*, 1653–1666.

[69] E. Vangrevelinghe, K. Zimmermann, J. Schoepfer, R. Portmann, D. Fabbro, P. Furet, *J. Med. Chem.* **2003**, *46*, 2656–2662.

[70] S. M. Rodems, B. D. Hamman, C. Lin, J. Zhao, S. Shah, D. Heidary, L. Makings, J. H. Stack, B. A. Pollok, *Assay Drug Dev. Technol.* **2002**, *1*, 9–19.

[71] *HotSpot*™, http://reactionbiology.com/pages/kinase.htm accessed on Sept. 15, 2009.

Supporting Information

High-throughput Virtual Screening using Quantum Mechanical Probes: Discovery of Selective Kinase Inhibitors

*Ting Zhou and Amedeo Caflisch**
Department of Biochemistry, University of Zürich,
Winterthurerstrasse 190, CH-8057
Zürich, Switzerland

Assessment of QM probe method on protein kinase CDK2

Before applying the quantum mechanical (QM) probe method to EphB4, it was tested on the cyclin-dependent kinase 2 (CDK2) for which a large number of ATP-binding site inhibitors have been published.[1–8] The structure of CDK2 was downloaded from the PDB (PDB entry 1KE5), and hydrogen atoms were added by CHARMM[9] according to the protonation states of side chains and termini at pH 7. Then the structure was minimized with CHARMM using the CHARMm[10] force field and MPEOE partial charges.

The catalytic domain in protein kinases is composed of two lobes connected by a segment termed "hinge loop". The majority of ATP-competitive inhibitors are involved in at least one hydrogen bond with the hinge loop.[11] There are two hydrogen bond (HB) acceptors and one donor in the backbone of the hinge loop so that two water probes and one hydrogen fluoride (HF), respectively, were used (Figure S-I). About 1,000 compounds, randomly selected from the ZINC library, were docked into the ATP-binding site of cyclin-dependent kinase 2 (CDK2) using version 4 of AutoDock[12]. A total of about 100,000 poses were generated by docking, then minimized using the CHARMm force field, and filtered by van der Waals (vdW) interaction energy (IE) and vdW efficiency as mentioned in ref 7. To study the effectiveness of the probe method, we selected poses based on probe energies and visually inspected whether there is a particular interaction at the expected position.

Figure S-II shows the QM probe energy is an effective detector of canonical HBs. The structures of four putatively inactive compounds (compound **27**–**30** termed decoys hereafter) and the cocrystallized ligand in PDB entry 1KE5, and their interactions with the protein are schematically shown in Figure S-II. From the vdW point of view, the four decoys match the binding pocket, since they all passed the filters of vdW and vdW efficiency.[7] However their unfavorable polar interactions with the hinge region are not detected. A main reason of failure in detecting the unfavorable polar contacts is that the E_{ele} averages out the electrostatic interaction

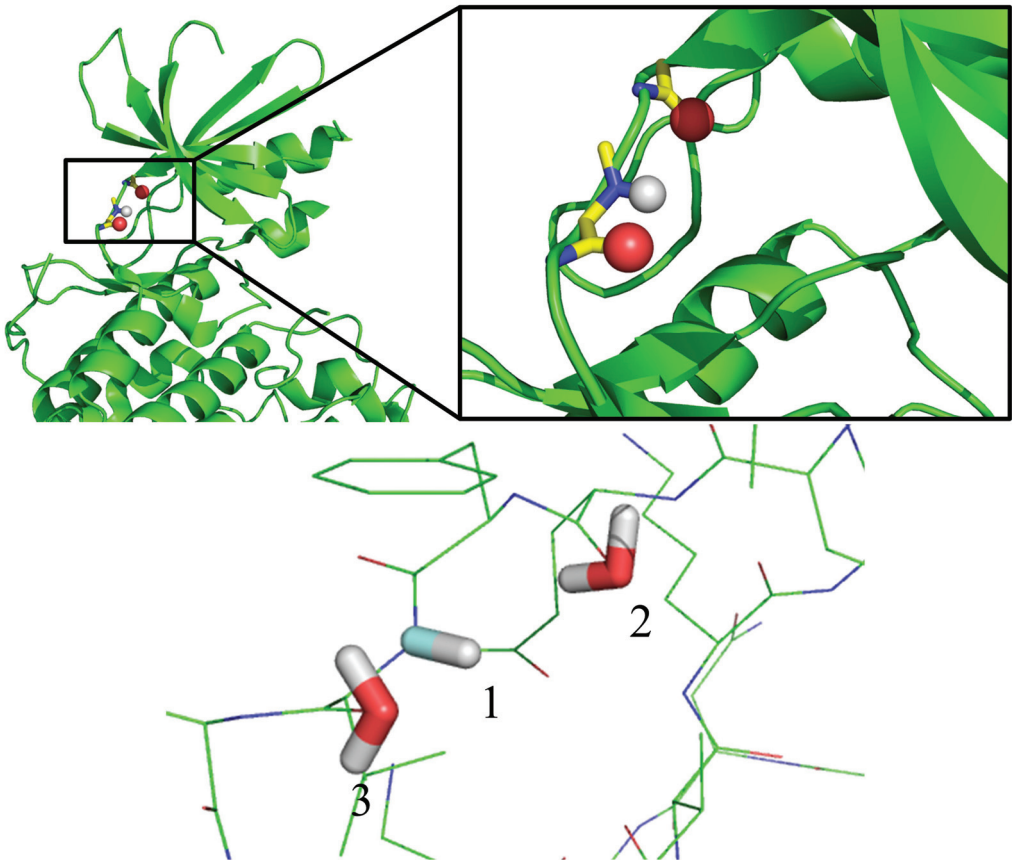
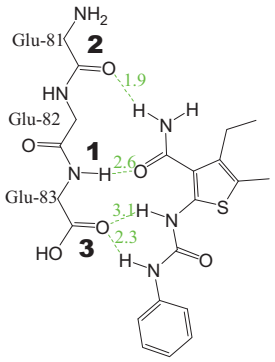


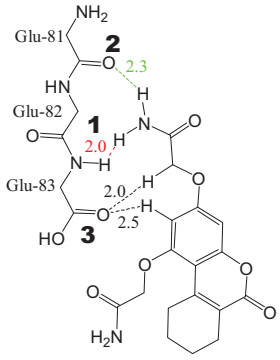
Figure S-I: The three QM probes at the hinge loop of the protein kinase CDK2. Positions 1, 2, and 3 are the backbone –NH– of Leu83, –CO– of Glu81, and –CO– of Leu83, respectively.

between the decoys and the protein, hence the resolution or sensitivity is not high enough. This averaging effect is also present for the known inhibitor (Figure S-II) in the structure of PDB entry 1KE5 which was optimized and evaluated using the same protocol, and it has $E_{\text{ele}} = -5.02$ kcal·mol^{-1}, and $E_{\text{ele}}/\text{MW} = -0.0153$ kcal·g^{-1}, where MW is the molecular weight. Electrostatic IE and electrostatic efficiency of this cocrystallized inhibitor are quite similar to some inactive compounds (Figure S-II). Nevertheless, with QM probe method, an adverse contact always shows a positive (unfavorable) probe energy, and can be easily distinguished. The three probe energies of the cocrystallized inhibitor are -3.94 kcal·mol^{-1}, -3.27 kcal·mol^{-1}, and -1.63 kcal·mol^{-1}, while each of the compound **28**–**30** has one positive probe energy, which is consistent with the unfavorable polar contact in the pose (Figure S-II). Note that compound **27** is not eliminated by the filters of the three probes at the hinge region, but was not selected for experimental testing, since the hydrophobic pocket is not satisfied (see subsection 3.3).

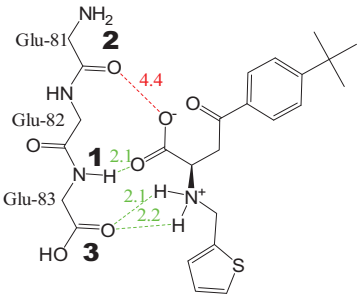
The QM probe method is able to detect non-classical HBs. Most of the force fields use fix-charge approximation to describe electrostatic interactions. QM gains an advantage in evaluation of complex charge–charge interactions, e.g., anion–cation interaction,[13] metal–ligand interaction,[14] and HB.[15] The non-classical HBs exist in protein–ligand complex extensively,[16] not only in kinase cases,[17] but also in other targets.[18] The compound in Figure S-II(b) is involved in a pair of C$-$H$\cdots$O non-classical HBs with the protein (colored in black).[17] The probe energy of Probe 3 is -3.22 kcal·mol^{-1}, which expresses a distinct sign of a favorable interaction there. Figure S-III shows another compound (compound **31**) interacting with the hinge loop at Probe 2 by a non-classical HB, whose probe energy (-2.02 kcal·mol^{-1}) is also comparable with that of a classical HB. The partial charge of the hydrogen atom bonded to the C_6 (red in Figure S-III) is identical with that of the hydrogen atom in unsubstituted phenyl ring if the partial charges are assigned using MPEOE approach. This is not accurate because the $-NO_2$ group is strongly electron-withdrawing, and the =CH= at the para position of the $-NO_2$ is more positive (atomic charges calculated by QM at MP2 6-31+G(d,p) level is shown in Figure S-VI) than the analogue



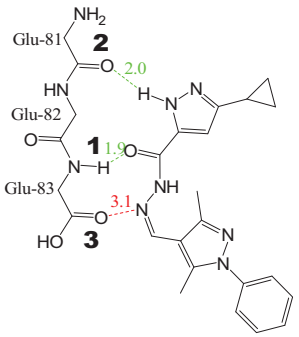
(a) Three probes match the decoy well.



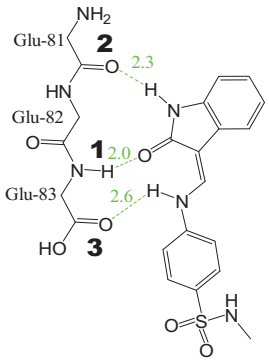
(b) Probe 1 mismatches the decoy.



(c) Probe 2 mismatches the decoy.



(d) Probe 3 mismatches the decoy.



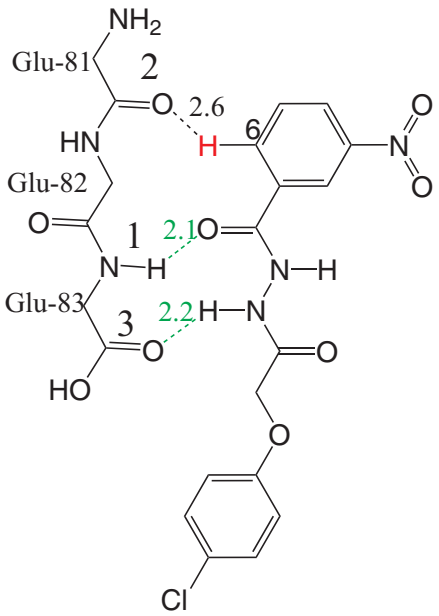
(e) Ligand in the structure of PDB entry 1KE5

Compound	ZINC ID	E_{ele}[a]	E_{vdW}	E_{ele}/MW	E_{vdW}/MW	Probe 1	Probe 2	Probe 3
27	476434	−1.6	−40.1	−0.005	−0.132	−2.38	−2.36	−3.65
28	43974	−4.0	−42.5	−0.012	−0.123	2.67	−3.43	−3.11
29	49031	−5.7	−35.5	−0.016	−0.103	−5.61	3.41	−7.22
30	55643	−0.9	−44.3	−0.003	−0.127	−4.63	−3.22	4.32
1KE5 Ligand		−5.0	−41.5	−0.015	−0.126	−3.94	−3.27	−1.63

(f) Probe energy of compounds.[b]

[a] Electrostatic interaction energy. [b] All energy values are in kcal·mol^{-1}. MW is in g·mol^{-1}.

Figure S-II: Assessment of QM probe method on CDK2 using four putatively inactive compounds (a)–(d), and a co-crystalized ligand (e). The distances between the critical atoms are noted with the digits above the dashed lines. The unit of length is Å. The green color denotes favorable HB interactions, the red indicates unfavorable interactions, and the black means favorable interactions but forming non-classical HBs.



31

Compound	ZINC ID	E_{ele}	E_{vdW}	E_{ele}/MW	E_{vdW}/MW	Probe 1	Probe 2	Probe 3
31	298885	-2.1	-39.5	-0.006	-0.113	-3.36	-2.02	-4.26

Figure S-III: Non-classical HB. All energy values are in kcal·mol^{-1}. The colors of the dashed lines and the digits have the same meanings as in the Figure S-II. The hydrogen atom in red is discussed in the main text.

without a $-NO_2$ group. Note that the experimental pK_a of nitrobenzene is 3.98 (at 0 °C).[19] Therefore, this hydrogen atom becomes a potential HB donor, and will form a HB when there is a HB acceptor nearby gaining a favorable interaction.

References

[1] L. Meijer, A. Borgne, O. Mulner, J. P. J. Chong, J. J. Blow, N. Inagaki, M. Inagaki, J. G. Delcros, J. P. Moulinoux, *Eur. J. Biochem.* **1997**, *243*, 527–536.

[2] S. H. Kim, *Pure Appl. Chem.* **1998**, *70*, 555–565.

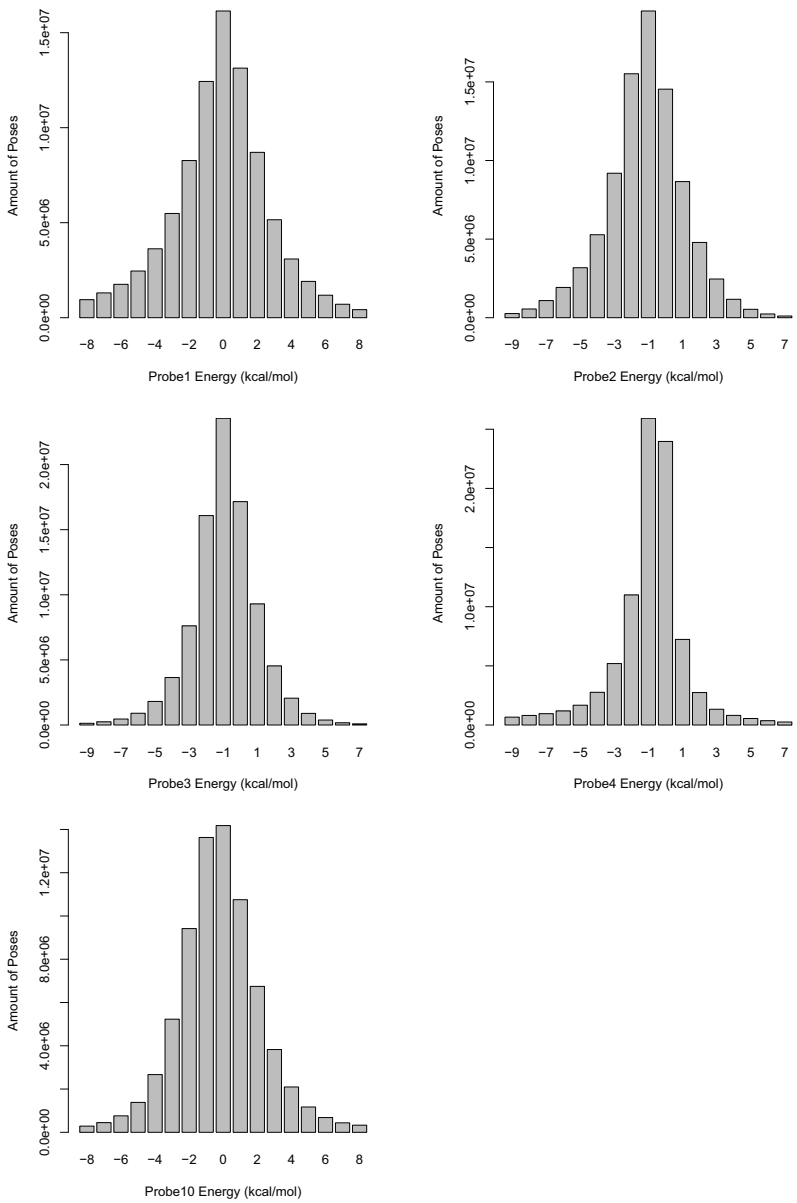
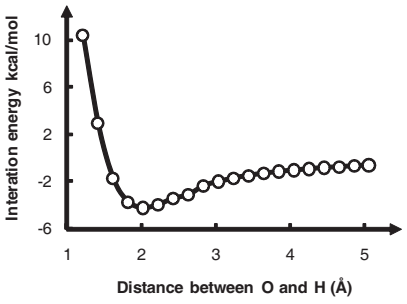
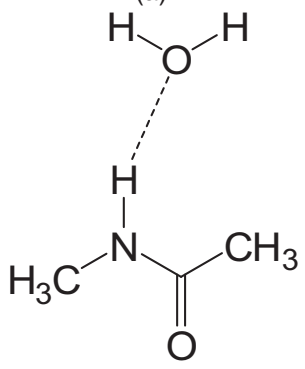


Figure S-IV: Distribution of energies of 5 probes (Probe 1–4, and 10) across 89,350,018 poses of neutral molecules.



(a)



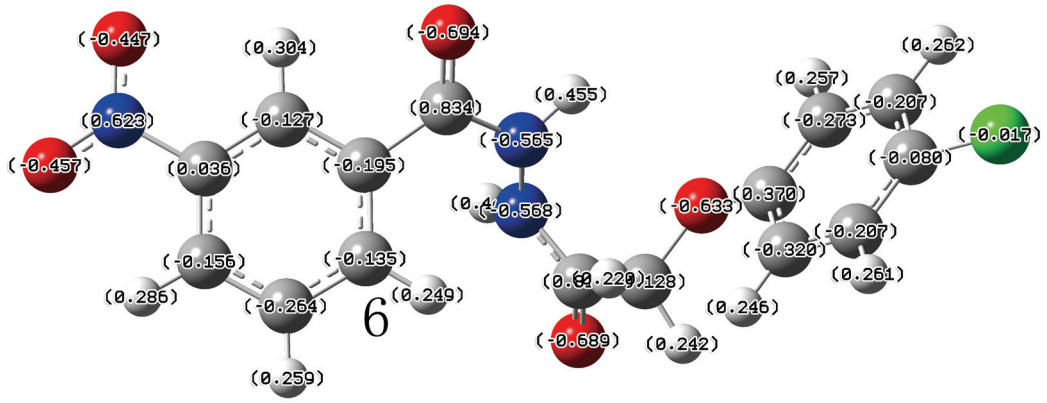
(b)

Distance(Å)	PM6	PM3	RM1	AM1
1.22	10.44	25.07	11.93	27.76
1.42	2.92	14.02	4.80	16.33
1.62	−1.74	1.92	0.38	5.69
1.83	−3.75	−1.97	−0.73	−1.09
2.03	−4.25	−1.18	−0.82	−3.53
2.23	−3.98	−1.02	−1.39	−3.73
2.43	−3.45	−1.45	−2.03	−3.24
2.64	−3.15	−1.52	−1.97	−2.74
2.84	−2.39	−1.41	−1.76	−2.43
3.04	−2.02	−1.26	−1.53	−1.70
3.24	−1.72	−1.11	−1.33	−1.39
3.45	−1.50	−0.97	−1.17	−1.17
3.65	−1.32	−0.86	−1.04	−1.02
3.85	−1.17	−0.77	−0.92	−0.90
4.06	−1.05	−0.69	−0.83	−0.81
4.26	−0.94	−0.62	−0.75	−0.73
4.46	−0.85	−0.56	−0.67	−0.66
4.66	−0.77	−0.51	−0.61	−0.60
4.87	−0.69	−0.46	−0.55	−0.55
5.07	−0.63	−0.42	−0.50	−0.50

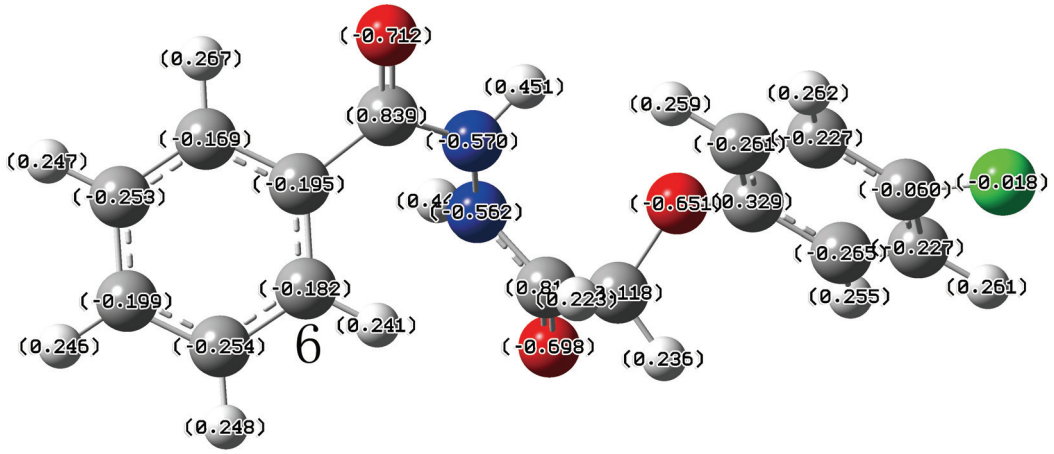
(c)

Figure S-V: (a) The PM6 interaction energies[a] between a water molecule and a N-methylacetamide are plotted against the distances from atom H to O connected with the dashed line in (b). The movement of the water molecule is along the vector defined by the atom H and O in the fully PM6-optimized conformation, i.e., the conformation when the distance equals to 2.03 Å in the second column of (c). After moving the water molecule to a new position, the conformation is partially optimized. The coordinates of atoms except for the H and the O atom are optimized using four Hamiltonians (PM6, PM3, RM1, and AM1) in MOPAC. The unit of all energy terms is kcal·mol^{-1}. The energy terms do not contain basis set superposition error correction.

[a] IE $= H_{\text{complex}} - H_{\text{water}} - H_{\text{N-methylacetamide}}$, where H is the formation enthalpy.



(1) Compound **31**



(2) Analogue of compound **31** without $-NO_2$ group

Figure S-VI: The quantum mechanical atomic charges of compound **31** and its analogue without $-NO_2$ group. The structures were minimized by B3LYP 6-31+G(d,p) starting with the docking structures. The digits in the parenthesis are the partial charges calculated by natural bond orbital theory at MP2 6-31+G(d,p) level. The group charge of C_6H changes from 0.059 electronic unit to 0.114 electronic unit, when a $-NO_2$ group substitutes its para hydrogen.

Table S-I: Probe energies of minor-different conformers of four known inhibitors of EphB4. The conformer in the first row of each block is the minimized conformer of each inhibitors which is identical to those listed in Table 2 of the main text. The other conformers are snapshots taken every 10 fs from a short molecular dynamics run (100 fs at 50 K) without minimization.

Conformer	Probe 1	Probe 2	Probe 3	Probe 4	Probe 10	RMSD of Coordinates(Å)
ALTA	−4.31	0.37	0.14	0.34	0.41	0.000
ALTA_1	−4.27	0.33	−0.04	0.33	0.34	0.025
ALTA_2	−4.32	0.34	0.20	0.35	0.34	0.026
ALTA_3	−4.16	0.39	0.05	0.30	0.43	0.030
ALTA_4	−4.18	0.34	0.25	0.36	0.38	0.031
ALTA_5	−4.24	0.34	0.08	0.30	0.40	0.035
ALTA_6	−4.39	0.40	0.05	0.37	0.43	0.031
ALTA_7	−4.18	0.38	−0.03	0.33	0.42	0.030
ALTA_8	−4.23	0.39	−0.03	0.35	0.41	0.031
ALTA_9	−4.16	0.39	0.31	0.35	0.46	0.034
ALTA_10	−4.12	0.35	−0.19	0.28	0.39	0.034
MIYA9f	−3.04	−2.25	−1.16	−5.53	−3.47	0.000
MIYA9f_1	−3.05	−2.17	−1.27	−5.49	−3.43	0.020
MIYA9f_2	−3.03	−2.26	−1.18	−5.45	−3.31	0.026
MIYA9f_3	−3.02	−2.39	−1.14	−5.64	−3.62	0.028
MIYA9f_4	−3.03	−1.96	−1.21	−5.27	−3.21	0.029
MIYA9f_5	−2.97	−2.25	−1.16	−5.50	−3.50	0.029
MIYA9f_6	−3.01	−2.22	−1.14	−5.25	−3.40	0.027
MIYA9f_7	−2.97	−2.09	−1.16	−5.35	−3.41	0.030
MIYA9f_8	−3.12	−2.15	−1.23	−5.52	−3.51	0.028
MIYA9f_9	−3.03	−2.03	−1.07	−5.15	−3.42	0.030
MIYA9f_10	−3.02	−1.99	−1.34	−5.23	−3.38	0.036
ONC102	−2.46	−2.43	−1.38	−0.22	−2.32	0.000
ONC102_1	−2.36	−2.40	−1.43	−0.20	−2.25	0.023
ONC102_2	−2.54	−2.43	−1.36	−0.22	−2.24	0.025
ONC102_3	−2.55	−2.49	−1.41	−0.20	−2.36	0.026
ONC102_4	−2.57	−2.54	−1.54	−0.22	−2.47	0.032
ONC102_5	−2.56	−2.34	−1.35	−0.20	−2.27	0.036
ONC102_6	−2.61	−2.38	−1.43	−0.21	−2.31	0.038
ONC102_7	−2.50	−2.33	−1.49	−0.21	−2.28	0.040
ONC102_8	−2.60	−2.33	−1.53	−0.20	−2.32	0.036
ONC102_9	−2.54	−2.50	−1.47	−0.22	−2.44	0.031
ONC102_10	−2.59	−2.51	−1.57	−0.21	−2.36	0.036
PP2	−3.71	−3.18	−1.24	0.12	−2.98	0.000
PP2_1	−3.73	−3.13	−1.34	0.12	−2.88	0.021
PP2_2	−3.62	−3.15	−1.26	0.12	−3.00	0.026
PP2_3	−3.72	−3.34	−1.24	0.11	−2.92	0.028
PP2_4	−3.63	−3.08	−1.23	0.15	−2.97	0.029
PP2_5	−3.63	−3.25	−1.09	0.14	−3.10	0.028
PP2_6	−3.57	−3.20	−1.21	0.09	−2.91	0.033
PP2_7	−3.71	−3.15	−1.13	0.15	−3.12	0.032
PP2_8	−3.65	−3.19	−1.17	0.10	−2.94	0.029
PP2_9	−3.60	−3.41	−1.06	0.13	−3.13	0.033
PP2_10	−3.65	−3.34	−1.14	0.08	−2.97	0.036

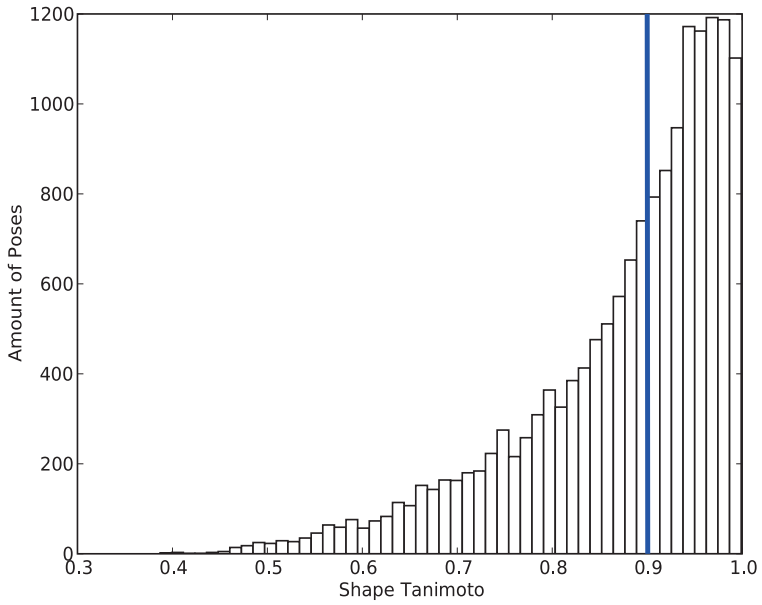


Figure S-VII: The distribution of shape Tanimoto of 15,979 poses. The blue vertical line at 0.9 emphasizes the threshold for filtering out the large-strain ligands.

Table S-II: List of 85 kinases for selectivity profile. The ATP concentration in each assay is denoted in the head row.

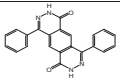
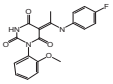
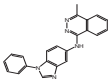
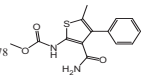
5μM*	20μM*	50μM*
ΔPH-PKBα (S473D)	Aurora B	ΔPH-PKBß (S474D)
CK2α	CaMKKß	AMPK
DYRK3	CDK2/cyclin A	BRSK2
EF2K	CHK1	BTK
EPH-B3	CHK2	CaMK1
ERK1	CK1δ	DYRK1a
ERK8	CSK	DYRK2
GSK3ß	FGF-R1	EPH-A2
HER4	GCK	IKKε
HIPK2	IR-HIS	LCK
IGF1R	IRAK4	MAPK2/ERK2
IKKß	JNK1α1	MAPKAP-K1a/RSK1
IRR	JNK2	MAPKAP-K1b/RSK2
MARK3	LKB1	MELK
MKK1	MAPKAP-K2	MINK1
p38γ MAPK	MLK1	MNK1
p38δ MAPK	MLK3	MNK2α
PAK4	MSK1	NEK2a
PIM2	MST2	NEK6
PKCζ	MST4	p38α MAPK
PLK1	NUAK1	PhKγ1
PRK2	p38ß MAPK	PKD1
	PAK5	smMLCK
	PAK6	Src
	PDK1	SRPK-1
	PIM1	TBK1
	PIM3	
	PKA	
	PKCα	
	PRAK	
	ROCKII	
	S6K1 (T412E)	
	SGK1	
	SYK	
	TTK	
	VEG-FR	
	YES1	

* The ATP concentrations are at or below the calculated Km for ATP for that kinase.

Table S-III: The sequence information and the gatekeeper residues of 85 kinases.

Protein Kinase	Accession No.	GI	Gatekeeper
AMPK[26-268]	NM_006252	46877068	M
Aurora B [1-344]	NM_004217	83776600	L
BRSK2 [2-674]	AF533878	33187742	L
BTK [2-659]	NP_000052.1	4557377	T
CaMK1 [2-369]	NM_003656	4502553	M
CaMKKβ [1-541]	NM_153499	27437017	F
CDK2[4-286]	NM_001798	16936528	F
CHK1 [1-476]	AF016582	2367669	L
CHK2 [5-543]	NM_007194	6005850	L
CK1δ [1-294]	AB063114	14422451	M
CK2α [2-391]	NM_001895	4503095	F
CSK [1-450]	NM_004383	4758078	T
DYRK1a [1-499]	NM_130437.2	18765754	F
DYRK2 [3-528]	NM_003583	4503427	F
DYRK3 [1-588]	AY590695	46909167	F
EF2K [2-725]	AAH32665	21618568	E
EPH-A2 [591-976]	NM_004431	32967311	T
EPH-B3 [561-998]	NM_004443	17975768	T
ERK1 [2-379]	BC013992	15559271	Q
ERK2 [1-358]	X58712	53002	Q
ERK8 [2-544]	AY065978	19263187	F
FGF-R1 [400-820]	M34641	182530	V
GCK [2 - 812]	BC047865	28839779	M
GSK3β [2-420]	L33801	529237	L
HER4 [706 - 991]	NM_005235	4885215	T
HIPK2 [165-564]	AF326592	17225377	F
IGF1R [954-1367]	NM_000875	4557665	M
IKKβ [1-736]	XM_032491	20538863	M
IKKε [1-716]	NM_014002	7661946	M
IR [1001-1382]	NM_000208.2	119395736	M
IRAK4 [140-460]	BC013316.1	15426432	Y
IRR [944-1236]	NM_014215	31657140	M
JNK1α1 [1-384]	L26318	474901	M
JNK2α2 [1-424]	L31951	598183	M
LCK [2-509]	X03533	244791455	T
LKB1 [1-433]	NP_000446	4507271	M
MAPKAP-K2 [46-400]	NM_032960	32481209	M
MARK3 [2-729]	U64205	3089349	M
MELK [2-651]	NM_014791	7661974	L
MINK1[1-320]	NM_015716	7657335	M
MKK1 [1-393]	Z30163	456202	M
MLCK [475-838]	NM_005965	16950601	L
MLK1 [132 - 413]	NM_033141	52421790	M
MLK3 [96 - 386]	NM_002419	4505195	M
MNK1 [2-424]	AB000409	2077825	F
MNK2α [2-465]	AF237775	11023170	F
MSK1 [2-802]	AF074393	3411157	L
MST2 [2-491]	U60206	1477789	M
MST4 [1-416]	NM_016542	15011880	M
NEK2A [1-445]	NM_002497	4505373	M
NEK6 [8-313]	NM_014397	19923407	L
NUAK1 [2-660]	NM_014840	7662170	M
p38β MAPK [1-364]	Y14440	2326554	T
p38α MAPK [1-360]	L35264	603919	T
p38γ MAPK [1-367]	Y10487	1785656	T
p38δ MAPK [1-365]	Y10488	2266640	M
PAK4 [2-591]	O96013	12585288	M
PAK5 [2-719]	Q9P286	12585290	M
PAK6 [2-681]	Q9NQU5	23396789	M
PDK1 [52-556]	NM_002613	4505695	L
PhKγ1[2-297]	X80590	1147567	F
PIM1 [2-313]	NM_002648	4505811	L
PIM2 [2-334]	U77735	1750276	L
PIM3 [2-326]	Q86V86	215274221	L
PKA [2-351]	NM_002730	4506055	M
PKBβ (S474D) [120-481]	NM_001626	4502023	M
PKBα (S473D) [118-480]	BC000479	12653417	M
PKCα [1-672]	NM_002737	4506067	M
PKCζ [2-592]	NM_002744	52486327	I
PKD1 [2-912]	NM_002742	115529463	M
PLK1 [1-603]	NM_005030	21359873	L
PRAK [1-471]	AF032437	3133291	M
PRK2 [501-984]	S75548	914100	M
ROCKII [2-543]	U38481	1384133	M
RSK1 [1-735]	M99169	206772	L
RSK2 [2-740]	NM_004586	4759050	T
S6K1 (T412E) [1-421]	NM_003161	4506737	L
SGK1 (S422D) [60-431]	NM_005627	25168263	L
Src [2-533]	NM_005417.3	4885609	T
SRPK1 [2-654]	NM_003137	47419936	F
SYK [1-635]	AAH01645.1	12804475	M
TBK1 [1-729]	NM_013254	7019547	M
TTK [1 - 857]	NM_003318	23308722	M
VEGFR [784-1338]	NM_002019.3	156104876	V
YES1 [1-543]	NM_005433	4885661	T

Table S-IV: Energy values in kcal·mol^{-1} of experimentally tested compounds[a].

ZINC ID	Structure	Probe1	Probe2	Probe3	Probe4	Probe5	Probe6	Probe7	Probe8	Probe9	Probe10	Probe11	Probe12	Probe13	hydrophobic matching	% inihibition at 50 µM
842896		−3.01	−2.51	0.21	−0.37	−0.03	0.12	0.28	0.12	−0.05	−0.97	−6.76	−8.52	−0.33	−9.27	19, 8
2361207		−4.49	−2.08	0.25	0.09	0.03	−0.04	0.40	0.11	−0.22	−2.90	−5.82	−3.78	−0.61	−7.96	∼0
1406465		−3.84	−1.25	−2.25	−1.38	−0.15	0.07	0.29	0.06	0.05	−0.46	−3.96	−2.98	0.14	−5.02	0, 54
1213337		−2.94	−1.08	0.43	−4.04	−0.01	−0.08	0.43	0.25	0.06	−1.69	−11.19	2.57	−0.29	−8.19	42, 70
1053478		−3.90	−2.34	−2.03	−1.05	0.10	−0.08	0.35	0.18	−0.23	−1.17	−7.04	0.61	0.23	−5.65	-NA-
838240		−3.16	−1.46	−0.27	−1.90	−0.37	−0.14	0.24	0.13	−0.07	−2.05	−7.22	−2.08	−0.33	−5.43	-NA-

[a] We do not have energy values of the 23 compounds mentioned in the main text because most of them are derivatives of ZINC compounds, since the original compounds were not available. The compounds **1** to **26** in the main text are derivatives of ZINC compound 1053478.

[3] B. B. McConnell, F. J. Gregory, F. J. Stott, E. Hara, G. Peters, *Mol. Cell. Biol.* **1999**, *19*, 1981–1989.

[4] T. G. Davies, P. Tunnah, L. Meijer, D. Marko, G. Eisenbrand, J. A. Endicott, M. E. M. Noble, *Structure* **2001**, *9*, 389–397.

[5] M. Nesi, D. Borghi, M. G. Brasca, F. Florentini, P. Pevarello, *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3205–3208.

[6] C. M. Richardson, C. L. Nunns, D. S. Williamson, M. J. Parratt, P. Dokurno, R. Howes, J. Borgognoni, M. J. Drysdale, H. Finch, R. E. Hubbard, P. S. Jackson, P. Kierstan, G. Lentzen, J. D. Moore, J. B. Murray, H. Simmonite, A. E. Surgenor, C. J. Torrance, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 3880–3885.

[7] P. Kolb, D. Huang, F. Dey, A. Caflisch, *J. Med. Chem.* **2008**, *51*, 1179–1188.

[8] Y. H. Jiang, R. C. H. Zhao, C. M. Verfaillie, *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 10538–10543.

[9] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comput. Chem.* **2009**, *30*, 1545–1614.

[10] F. Momany, R. Rone, *J. Comput. Chem.* **1992**, *13*, 888–900.

[11] J. J. Liao, *J. Med. Chem.* **2007**, *50*, 409–424.

[12] D. S. Goodsell, A. J. Olson, *Proteins* **1990**, *8*, 195–202.

[13] T. Zhou, D. Huang, A. Caflisch, *J. Med. Chem.* **2008**, *51*, 4280–4288.

[14] K. Raha, K. M. Merz, *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.

[15] J. Sponer, J. Leszczynski, P. Hobza, *Biopolymers* **2001**, *61*, 3–31.

[16] S. Sarkhel, G. R. Desiraju, *Proteins* **2004**, *54*, 247–259.

[17] A. C. Pierce, K. L. Sandretto, G. W. Bemis, *Proteins* **2002**, *49*, 567–576.

[18] B. Klaholz, D. Moras, *Structure* **2002**, *10*, 1197–1204.

[19] E. P. Serjeant, B. Dempsey, *Ionization Constants of Organic Acids in Aqueous Solution*, Pergamon, Oxford, **1979**.

Chapter 4

Data Management System for Distributed Virtual Screening

Zhou, T.; Caflisch A. *J. Chem. Inf. Model.* **2009**, *49 (1)*, 145–152

Data Management System for Distributed Virtual Screening

Ting Zhou and Amedeo Caflisch*

Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received August 22, 2008

High throughput docking (HTD) using high performance computing platforms is a multidisciplinary challenge. To handle HTD data effectively and efficiently, we have developed a distributed virtual screening data management system (DVSDMS) in which the data handling and the distribution of jobs are realized by the open-source structured query language database software MySQL. The essential concept of DVSDMS is the separation of the data management from the docking and ranking applications. DVSDMS can be used to dock millions of molecules effectively, monitor the process in real time, analyze docking results promptly, and process up to 10^8 poses by energy ranking techniques. In an HTD campaign to identify kinase inhibitors a low cost Linux PC has allowed DVSDMS to efficiently assign the workload to more than 500 computing clients. Notably, in a stress test of DVSDMS that emulated a large number of clients, about 60 molecules per second were distributed to the clients for docking, which indicates that DVSDMS can run efficiently on very large compute cluster (up to about 40000 cores).

INTRODUCTION

In silico screening of large libraries of compounds is a commonly used tool in drug discovery because it efficiently identifies candidate lead compounds.[1−7] Its efficiency is due to methodological progresses and the ever increasing performance of ordinary low-cost computers. Despite these progresses, handling large libraries of compounds is still a challenge for data management in drug design and discovery. The demand for an efficient data management increases even further if multiple computing instances access and update the data simultaneously.

Recently, several applications of large-scale virtual screening in parallel have been reported. For instance, para_glide, a recently developed extension of Glide[8,9] for parallel execution, counts the number of ligands, divides them into equal segments, and distributes them over several processors or machines. At the end of all docking calculations it provides a unified report of docking scores. Zhang and co-workers have developed the free package DOVIS which runs AutoDock[10] in parallel.[11] With DOVIS users can submit multiple jobs from a graphical user interface to both cluster and standalone computers. The authors docked about 2 million compounds on a Linux cluster with 256 CPUs and observed near-optimal performance. The essential concept of para_glide and DOVIS is the splitting of the molecular database into multiple partitions, which are then submitted to different processors individually, and the results are retrieved from all processors and combined after docking. In this way, the number of partitions and the amount of molecules are determined before splitting, the running time of each partition cannot be estimated, and the balance of each processor cannot be guaranteed either. Moustakas and Kuntz developed the MPI version of DOCK,[12] which used a master-worker scheme for parallelization.[13] To reduce bookkeeping tasks associated with manual partition of jobs, data were distributed to workers as molecules were read by the master, such that the poor load balance due to the random distribution of jobs was circumvented. Furthermore, Peters and co-workers recently optimized and validated DOCK on a massively parallel system with more than 16000 processors.[14] They pointed out that as the number of processors increased, the HTC (High Throughput Computing)[15] version of the DOCK program was more efficient than the MPI version, since library docking could be run as a collection of independent tasks while the MPI version suffered from overloading of the master. In other words, the efficiency of distribution of the master is the bottleneck of the master-worker scheme, in particular when a considerable amount of workers request jobs simultaneously. As an example, the efficiency of the MPI version of DOCK is 88% at 8192 workers but decreases to 55% at 16384 workers.[14] The overloading has been overcome by employing multilevel master-worker scheme (MLMW). However, both HTC and MLMW require additional time-consuming programming, in particular, HTC demands for the implementation of an asynchronous task dispatch subsystem, while MLMW requires the modification of the source code of the docking software.

The efficiency of data management is crucial in parallel applications. Furthermore, there is a strong demand of an efficient and easy-to-implement procedure to handle the data for a large number of computing clients. At present, most docking software reads the input and stores the output in plain files directly. Nonetheless, storing massive data in plain files is not suitable for extensive data management, since it usually requires more application programming effort to create, modify, and access data efficiently and securely. A database management system is a computer software designed to handle massive data efficiently. Providing controls of communication and synchronization, it allows multiple tasks to access and update the data in parallel with marginal

* Corresponding author phone: (+41 44) 635 55 21; fax: (+41 44) 635 68 62; e-mail: caflisch@bioc.uzh.ch.

10.1021/ci800295q CCC: $40.75
Published on Web 12/12/2008

Table 1. Structure of Table "ZINCMOL"[a]

column name	data type	explanation
zincmol.id	int(11) unsigned not null auto_increment	a unique identity for each molecule
zincmol.numatoms	tinyint(3) unsigned not null	number of atoms
zincmol.numc	tinyint(3) unsigned not null	number of carbon atoms
zincmol.numn	tinyint(3) unsigned not null	number of nitrogen atoms
zincmol.numo	tinyint(3) unsigned not null	number of oxygen atoms
zincmol.numhal	tinyint(3) unsigned not null	number of halogen atoms
zincmol.nums	tinyint(3) unsigned not null	number of sulfur atoms
zincmol.nump	tinyint(3) unsigned not null	number of phosphorus atoms
zincmol.numarombnd	tinyint(3) unsigned not null	number of aromatic bonds
zincmol.numdoubbnd	tinyint(3) unsigned not null	number of double bonds
zincmol.numtribnd	tinyint(3) unsigned not null	number of triple bonds
zincmol.numamibnd	tinyint(3) unsigned not null	number of amide bonds
zincmol.numacc	tinyint(3) unsigned not null	number of hydrogen bond acceptors
zincmol.numdon	tinyint(3) unsigned not null	number of hydrogen bond donors
zincmol.numring	tinyint(3) unsigned not null	number of rings
zincmol.totringsize	tinyint(3) unsigned not null	number of heavy atoms in rings
zincmol.longestchain	tinyint(3) unsigned not null	longest chain of atoms in the molecule
zincmol.wienerind4	float(16,14) not null	Wiener index
zincmol.numbnd	tinyint(3) unsigned not null	number of bonds
zincmol.numfrg	tinyint(3) unsigned not null	number of fragments
zincmol.numrotbnd	tinyint(3) unsigned not null	number of rotatable bonds
zincmol.mw	float(8,3) unsigned not null	molecule weight
zincmol.clogp	float(5,2) not null	CLogP
zincmol.charge	int(2) not null	formal charge
zincmol.mol2file	blob	compressed mol2 file
zincmol.tag1	char(20) default null	notes of calculation status for first target
zincmol.tag2	char(20) default null	notes of calculation status for second target
...	...	...
zincmol.tag*n*	char(20) default null	notes of calculation status for *n*th target

[a] The data types are represented in MySQL syntax.[33] The column "zincmol.id" is the primary key. The auto-incremental identifier can be used to discriminate individual protonation states and/or tautomeric forms. The 23 following columns contain the atomic and chemical properties of a molecule. The column "zincmol.mol2file" contains the compressed molecule file in mol2 format. The last columns "zincmol.tag*n*" (tag columns) record the calculation status for each protein target (e.g., multiple structures of protein or multiple proteins). Besides the primary key on column "zincmol.id", indexes are built on tag columns to speed up checking of the status by the computing clients. Other columns were not indexed because there was no query on them in the applications presented here.

additional effort in programming. It contains mature facilities to keep data integrated and consistent and provides utilities for database maintaining, such as backup, recovery, monitoring, and tuning.

In this paper, we introduce the Distributed Virtual Screening Data Management System (DVSDMS) for docking and ranking based on a master-worker scheme and the freely available database software MySQL. By separating operations related to data management from the main application and allocating the former to the database management system, DVSDMS can manage HTD data effectively and efficiently. The connection between the different programs is handled via scripts written in Python. A MySQL database, the master of the system, is used for organizing all the data. The clients (workers) do not communicate among each other but only with the database. In the application presented here, AutoDock was used as the docking engine, while DAIM,[16] Witnotp,[17] CHARMM,[18] and MOPAC[19] were employed for preparing the compound library, file type conversion, minimizing poses, and quantum mechanical calculations, respectively. Because of the flexibility of DVSDMS, other programs can be used for docking and ranking. Alternative procedures range from simple molecular mechanics energy terms and combinations thereof[20] to a quantum mechanics based ranking approach.[21] Furthermore, it is straightforward to use DVSDMS for consensus scoring.[22−24] In fact energy values calculated by different scoring functions can be stored in tables for ranking.

DVSDMS was validated in this study by docking about 1.5 million compounds into the ATP-binding site of the receptor tyrosine kinase EphB4 and ranking about 100 million poses using two Beowulf clusters (located on the same grid but with a number of switches ranging between 1 and 4). In the productive phase of docking and ranking the average load of the database management system was less than 10% and 30% with more than 500 workers requesting jobs, respectively. In a stress test, the database server built on a low cost Linux PC was able to distribute about 60 molecules per second.

DOCKING AND RANKING BY DVSDMS

The following three subsections describe briefly the overall process and programs used for docking and ranking in this application of DVSDMS. Details of the DVSDMS architecture are given in the next section.

Predocking. All structures and properties of the molecules required for docking and ranking were calculated and stored in the database. For each molecule in the ZINC library[25] (version 7) CHARMm[26] atom types were assigned with Witnotp.[17] Then DAIM[16] was applied to calculate the atomic and chemical properties of each molecule (listed in Table 1). Even though not all of these properties were used in docking and ranking, they were prepared for different kinds of filters one might want to apply before docking. Besides these properties, the mol2 file of each molecule was also

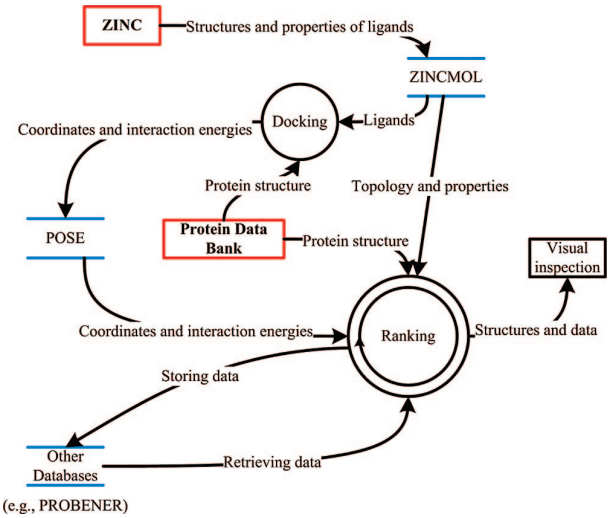


Figure 1. Schematic representation of docking and ranking processes. Red boxes indicate the public-domain databases, while blue parallels mean the tables in DVSDMS.

Table 2. Structure of Table "POSE"[a]

column name	data type	explanation
pose.id	int(11) not null auto_increment	a unique identity for each pose
pose.ele	float	electrostatic interaction
pose.elee	float	electrostatic efficiency
pose.vdw	float	vdW interaction energy
pose.vdwe	float	vdW efficiency
pose.pdbfile	blob	compressed pdb file
pose.mol_id	int(11) unsigned not null	related zincmol.id in Table "ZINCMOL"

[a] The data types are represented in MySQL syntax.[33] The Column "pose.id" is the primary key. An index is built on Column "pose.mol_id", which is a pointer for connecting the record in Table "POSE" to the one in Table "ZINCMOL". The value of Column "pose.mol_id" equals to the value of the primary key of Table "ZINCMOL".

stored in the database. A table termed "ZINCMOL" (Table 1) was designed to store these data (Figure 1).

Docking. AutoDock[10] (version 4) was applied for docking small molecules from the library into the receptor (see the Supporting Information). The poses of each molecule in the PDB format, with their interaction energies with the receptor and efficiencies (electrostatics and vdW), were stored in the table "POSE" (Table 2) of the database. During the docking, the computing clients acquired the 3D structure of the molecules directly from the database and stored poses and energies in the database after each docking process finished (Figure 1).

Ranking. Different scoring approaches can be handled efficiently by DVSDMS. In the present application to the EphB4 kinase, we used an in-house developed approach based on calculations of semiempirical quantum mechanics to efficiently rank the poses (Zhou et al., manuscript in preparation). Ranking a pose was usually faster than docking a molecule; therefore, the former needed a more efficient database I/O environment than the latter (see Results).

ARCHITECTURE OF DVSDMS

Database Structure. All the actions of the computing clients were coordinated by the database server (Figure 2). Three tables are necessary in the database for handling poses and molecules: one for the data of the molecules (Table 1), one for the poses from docking (Table 2), and one for the results of the ranking process (Table 3). Besides the properties of the molecules and mol2 files, the column "zincmol.tag" was added to Table "ZINCMOL" to record the status of docking. The format of the "zincmol.tag" column is shown in Figure 3. In the case of multiple protein targets, additional tag columns can be appended for each of the targets to record the status of docking processes. If the amount of targets is large and the appended tag columns affect the efficiency of the database, the ZINCMOL table can be vertically partitioned like the POSE table (not used in this application, see Partitioning Large Tables). The main columns, such as the properties and the coordinates of molecules, do not need to be copied for each target. If the molecule has not been handled for the specific target, the related "zincmol.tag" column is set to "null". Before the docking starts the non-drug-like molecules (according to user-defined filters) can be marked as "not passed" at "zincmol.tag" columns (see Results). The database server returns one of the molecules with a "null" tag when a docking client requests a job. The molecules which were marked as "not passed" will not be returned to computing clients. The "pose.tag" column of Table "POSE" (Table 2) was partitioned into another table ("POSETAG", Table 4) to improve the database performance (see **Partitioning Large Tables**). Column "posetag.sign" in the Table "POSETAG" is the status of ranking process of the related pose: "1", "2", "3", and "6" mean "in process", "finished normally", "finished with errors", and "unhandled", respectively. When a ranking client requests a job, the database returns a set of poses with "unhandled" signs.

Interface between Software Packages. Python scripts were developed for connecting all software packages in DVSDMS. The whole process needed the cooperation of several types of software (Table 5) developed by different groups. Therefore, some jobs such as converting file type, preparing input files, and parsing output files were needed to connect each stage of the pipeline. The communication to the database is essential in the DVSDMS. The Python package SQLAlchemy[27] is used to establish the crucial connection of python scripts (connecting the different stages of the docking pipeline) and the MySQL database (the data and result storage facility). The package Elixir[28] (object-relational mapping features) facilitates the consequent treatment of all data entries as objects, which results in clean code such that raw SQL statements are used only in performance critical parts.

Running on Standalone and Cluster Computers. DVSDMS runs as a single executable Python script on standalone or cluster computers without a queuing system. In the presence of a queuing system DVSDMS can be submitted to the queue with a runtime limit in accordance with the configuration of the queuing system. In this case, the DVSDMS client estimates the execution time of the next job before acquiring it from the database.

Monitoring Process and Identifying Errors. Users can monitor progress of the computation and trace errors by means of "sign" and "tag" columns of the corresponding tables in DVSDMS. The docking status of a molecule, machine name of the client, and the starting time of the job

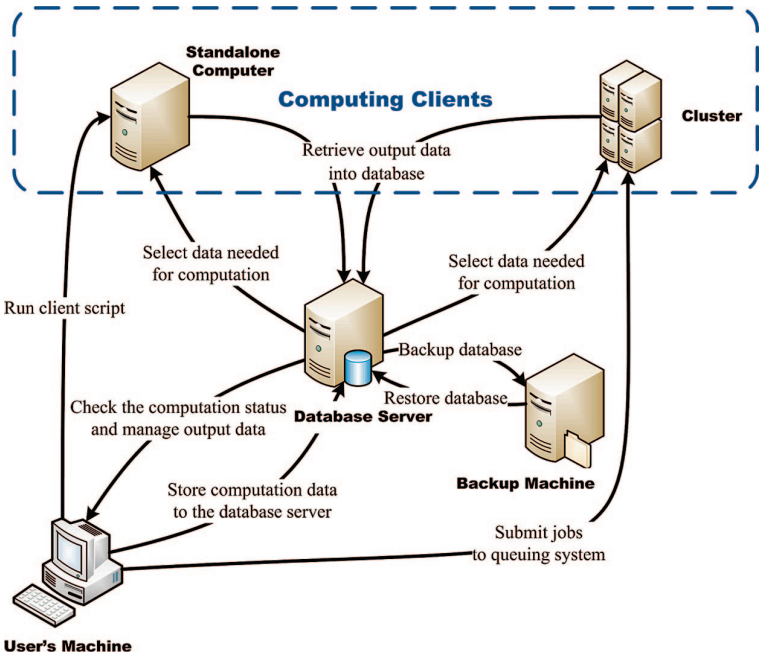


Figure 2. Hardware and data flow in DVSDMS. Note that in the application presented here the user's machine, backup machine, and database server were all on a single PC.

Table 3. Structure of Table for Ranking (PROBENER in Our Application)[a]

column name	data type	explanation
probener.id	int(11) not null auto_increment	auto increment "probener" id
probener.p*1*	float	interaction energy with the first probe
probener.p*2*	float	interaction energy with the second probe
...	...	...
probener.p*n*	float	interaction energy with the *n*th probe
probener.pose_id	int(11)	related pose.id in Table "POSE"

[a] The data types are represented in MySQL syntax.[33] The Column "probener.id" is the primary key. An index was built on Column "probener.pose_id", which is a pointer for connecting the record in Table "PROBENER" to the one in Table "POSE". The value of Column "probener.pose_id" equals to the value of the primary key of Table "POSE".

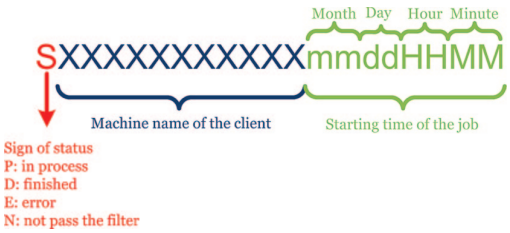


Figure 3. Format of "zincmol.tag*n*" column in Table "ZINCMOL".

can be read from Column "zincmol.tag" in Table "ZINCMOL" (Figure 3). In Table "POSETAG", Column "posetag.sign" is further separated from Column "posetag.tag" (Table 4) because in the I/O intensive ranking stage, the database only needs to scan "posetag.sign" to attain the status of the pose when requested for an "unhandled" pose.

Table 4. Structure of Table "POSETAG"[a]

column name	data type	explanation
posetag.id	int(11) not null auto_increment	auto increment posetag ID
posetag.pose_id	int(11)	related pose.id in Table POSE
posetag.sign	int(2) unsigned default '6'	sign of status
posetag.tag	char(20) default null	note for calculation status

[a] The data types are represented in MySQL syntax.[33] The Column "posetag.id" is the primary key. The Column "posetag.sign" was introduced for efficient retrieval of the ranking status. The meanings of signs are mentioned in the main text. An index was built on Column "posetag.pose_id", which is a pointer for connecting the record in Table "POSETAG" to the one in Table "POSE". The value of Column "posetag.pose_id" equals to the value of the primary key of Table "POSE".

Distributed computing systems are more prone to error than a standalone computer. It is very labor-consuming to reroll the process and locate errors when millions of compounds are handled in a high-throughput screening campaign. DVSDMS records stage information of clients in "sign" and "tag" columns and can check the status of jobs with user-defined frequency. Practically if a job does not finish during a given period of time, its status will be set to "unhandled", and its executing client will be reported. Then the job returns to the waiting list and is ready to be assigned to another free client.

PERFORMANCE TUNING

The optimization of DVSDMS focuses on the database performance tuning because, as mentioned above, the data management is separated from the main docking and ranking applications in DVSDMS. In the following, details on the process of optimization are given.

Table 5. Software and Its Function

software name	function	ref
DAIM	calculate the properties of molecules	16
AutoDockTools	convert molecule file into pdbqt type	34
AutoDock	dock small molecules into receptor	10
CHARMM	add hydrogen atoms, and minimize structure	18
Witnotp	convert molecule file types among mol2, pdb, and psf	17
MOPAC	calculate QM energies used for ranking	19
MySQL	the database management system	35
SQLAlchemy	the database toolkit for Python	27
Elixir	a declarative layer on top of SQLAlchemy	28

Using Database Index. An index was created on the column which was of high querying frequency. Each table has a primary key, a unique index to identify each row in a table, used for attaining the specific record promptly. For instance, with the primary key, any molecule record in the Table "ZINCMOL" can be retrieved out of millions of others by its ID stored in the Column "zincid.id" (Table 1) in one millisecond after the index was cached in the memory (see RESULTS section for hardware description). This is also valuable when the application needs to find out which molecule a given pose belongs to via Column "pose.mol_id" in Table "POSE" (Table 2). Similarly, but in an inverse way, the index of Column "pose.mol_id" (Table 2) is of use for fast reverse query, e.g., attaining poses related to a molecule. Only columns frequently queried are indexed. Other columns, e.g., interaction energies of poses and ranking scores, are not indexed because each additional index increases the size of the database and reduces the writing speed of tables.

Partitioning Large Tables. The performance of the database can be improved by partitioning large tables. During the database scan operation initiated by a query, only partitions containing the data are accessed, and during the maintenance, only damaged partitions instead of the entire table are repaired. Furthermore, the partitioned tables can be distributed on different physical drives, and tables can be scanned in parallel to improve both CPU and disk performance (which was not necessary for the present application). Two major forms of partitioning were applied in our database:

Horizontal Partitioning: Tables "ZINCMOL" and "POSE" were horizontally segmented into 50 partitions according to the hash function of their primary keys. Partitioning by hash is used primarily to ensure an even distribution of data among a predetermined number of partitions.[29] The value of a hash function determines the membership of a partition, e.g., the hash function returns an integer from 0 to 49 in the case with 50 partitions. The horizontal partitioning feature is supported by MySQL starting from version 5.1.

Vertical Partitioning: The most referenced columns "sign" and "tag" in the original Table "POSE" were separated into a new table "POSETAG" (Figure 4). Table "POSE" and "POSETAG" referred to each other via the column "pose.id" in "POSE" and the column "posetag.pose_id" in "POSETAG".

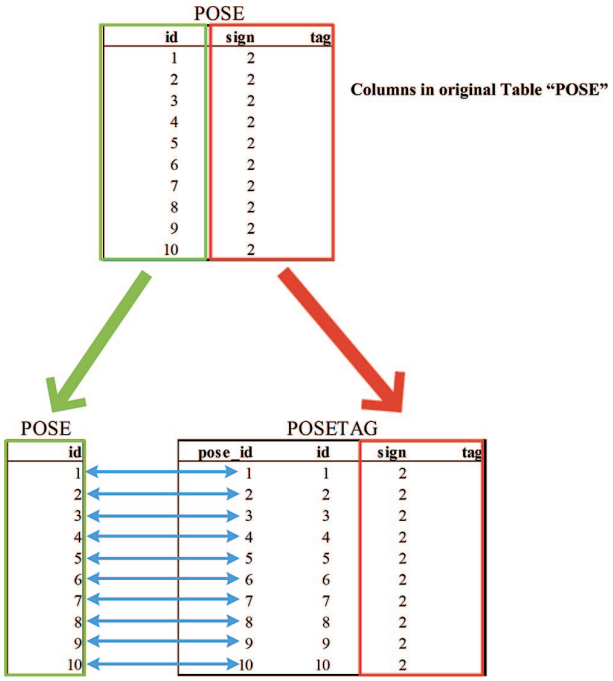


Figure 4. Vertical partitioning of Table "POSE". The value of Column "posetag.id" in "POSETAG" (Table 4) has no relation to that of Column "pose_id". The value of Column "sign" can be "1", "2", "3", and "6", which mean "in process", "normally finished", "finished with errors", and "unhandled", respectively.

This relation is analogous to foreign key constraints in the context of relational databases, even though MyISAM[30] (see Storage Engine) still does not support it in the latest version (MySQL 6.0). In this way, when handling the status of poses, the database management system only scanned the smaller table ("POSETAG") with the fixed row format instead of the large and dynamic table ("POSE"). In addition, different storage engines could be applied on different tables (see **Storage Engine**).

Storage Engine. Most tables in the database are constructed with MyISAM,[30] which is the default storage engine of MySQL due to its very low overhead, except for Table "POSETAG" constructed with InnoDB.[31] InnoDB uses more memory as cache to achieve a high performance. In fact, the database engine does not allow parallel accesses: a client has to lock the object for an update to prevent conflicting with other clients. InnoDB implements row-level locking, so that InnoDB only locks the rows needed for update instead of locking the entire table as MyISAM does. This feature is advantageous to concurrent updating from multiconnections with low lock wait ratio (LWR). The LWR is the percentage of queries that are required to wait for object locks to be released so that the query can itself acquire a lock on the object, e.g., many clients can update statuses of poses by modifying the Table "POSETAG" synchronously. The parallel performance of DVSDMS can be estimated by the LWR. A low LWR means that the performance loss due to the multiple connections of database is marginal. By using InnoDB instead of MyISAM as the storage engine of "POSETAG", the LWR of the database is reduced significantly, specifically in the ranking process MyISAM often induced a deadlock (LWR≈100%) while InnoDB reduced LWR to less than 0.1%.

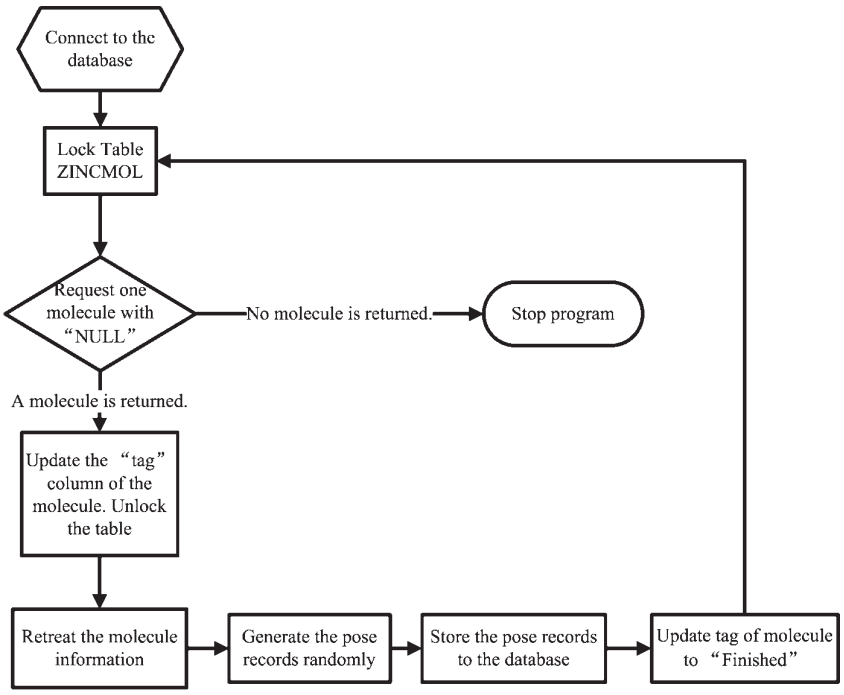


Figure 5. The flowchart of the docking emulator. In the benchmark the average amount of poses for each molecule is 37 ± 3 and the average size of each pose is 1 ± 0.5KB, which are consistent to the average in the real application to identify kinase inhibitors.

Local Cache and Bulk Update. The performance of the database can be improved by using local cache and bulk update, which were especially important for the short-term process, such as ranking in our calculation. Otherwise, the applications communicated with the database with high frequency and cast a heavy burden on the database. In our application, 2000 poses were retrieved from the database by a single SQL command and stored in the local memory. During the calculation, every result of single pose was stored in memory temporarily, which works as local cache. After all the ranking calculations of these poses had been finished, the client sent the results to the database and updated multiple rows (bulk update) by another SQL command. In this way, the clients only need to communicate with the database twice for handling 2000 poses. Note that if the amount of poses retrieved by the client using a single SQL command is too small, there will be no obvious increment of performance. Conversely, a large amount of poses will increase the individual query time and the memory use of the database server.

Compressing Molecule Files. Compressing molecule files is favorable for saving disk space and network bandwidth and decreasing the I/O intensity. For a normal PDB or MOL2 file, compression reduced its size by about 75% using zlib.[32] The compression and decompression work was performed by the clients and would not influence the load of the database server. After compression, the Table "POSE" used 93.1GB hard disk for about 100.8 million poses and their energies (Table 2).

RESULTS

Benchmark. Since we could not access a large computer cluster, we wrote an emulator of a docking client to test the performance of DVSDMS under conditions typical of very large clusters where the bottleneck is the master rather than the clients. In fact, our emulator does not run the real docking calculation but only requests jobs from the master and returns the output data, such as randomized interaction energies and efficiencies as well as binding poses (Figure 5). Note that emulated docking does not require any CPU time, which is essential to estimate the maximal rate of distribution of molecules (maxRoD) to the computer clusters we could access (≤200 nodes). The benchmark database server was built on a dedicated desktop computer, which had two Xeon 3.06 GHz CPUs, 2 GB of RAM, and a 250GB normal speed hard disk drive (maximal reading speed 50MB/s). The efficiency of a parallel docking program based on the master-worker scheme can be measured by the maxRoD of the master. A series of test cases was performed with the amount of emulators ranging from 1 to 160 and running one emulator per core. Two Beowulf clusters at the University of Zürich were used: The Etna and Matterhorn, which are separated from the master by one and four switches, respectively. Both clusters and the master are on a Gigabit network. The time required for emulated docking of 128000 molecules, which is a subset of ZINC library, is shown in Figure 6 for different numbers of emulators. The minimal times for the simulated docking are 1816 and 2292 s on the Etna and Matterhorn cluster, respectively. The lower amount of switches to reach Etna yields about 26.2% performance improvement in the benchmark. Note that the system load of the master does not hit 100% when the maxRoD reaches the plateau at about 12 cores (or emulators), which indicates that the network delay rather than the capacity of the master limits the overall performance. The maxRoD of DVSDMS running on Etna and Matterhorn are 70 and 56, respectively. The MaxRoD of the MPI version of DOCK[14] is about 19 (see the Supporting Information). Therefore the DVSDMS is about two times faster than the MPI version of DOCK on the Blue Gene/L platform. Even though the HTC version of DOCK

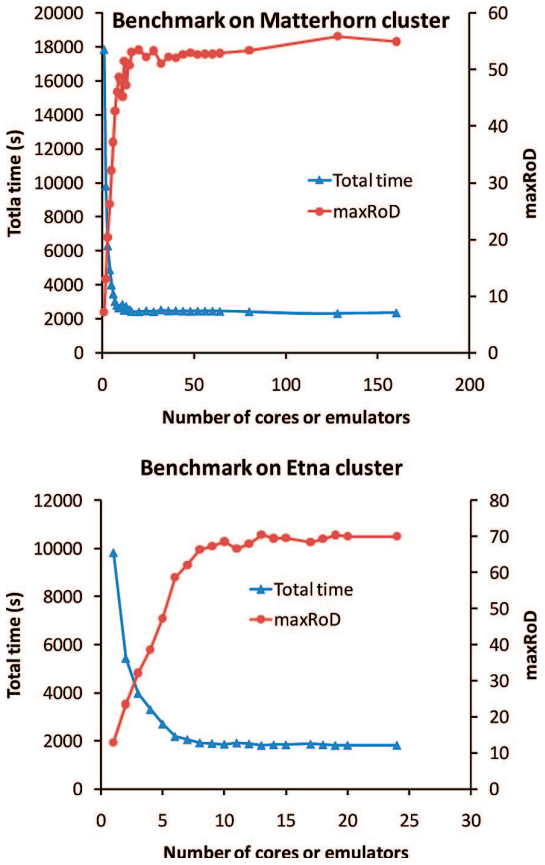


Figure 6. The duration of simulated docking of 128000 molecules with different numbers of emulated docking clients. Note that each core runs only one emulator. (Top) The emulators were submitted to the Matterhorn cluster (80 compute nodes each with dual-processor Opteron 2.4 GHz or 2.6 GHz), from which data needed to pass 4 switches to reach the database server. With more than 12 emulators, the database server, i.e., the master of DVSDMS, achieves a maximal rate of distribution of molecules (maxRoD) of about 56 per second on average (128000/2292s). (Bottom) On the Etna cluster (3 compute nodes each with dual Quad Core Xeon 2.33 GHz), the data only needed to pass 1 switch to reach the database server, and the maxRoD increased to about 70 per second on average (128000/1816s). Note that the delay due to the network equipment is negligible when there is a small quantity of requests. In both benchmarks however, the database server needs to respond to thousands of requests per second, so that the network delay limits the overall performance, and the plateau of the maxRoD is due to the network.

and the DVSDMS have similar capacities for distributing molecules (up to at least 16384 processors) it is easier to implement other docking engines in DVSDMS than to write a specific HTC version for each docking engine.

Performance in Production. The database server, backup machine, and the user's machine (Figure 2) were all built on the Linux desktop PC of the first author, which had dual-core 3.4 GHz Pentium D, 3 GB of RAM, and 3×320GB hard disks. The disk could read data at about 70 MB/s.

About 1.5 million compounds out of 3.8 millions in the ZINC library passed the filters used for eliminating the non-drug-like molecules (molecular weight <500 Da, number of rotatable bonds ≤7, number of hydrogen bond donors ≥1, and number of hydrogen bond acceptors ≥1). For the compounds which did not pass the filters the column "zincmol.tag" was set to "not passed". About 15 to 20 min were required for docking a single compound into the receptor and CHARMM minimization of the poses (with a rigid protein) on an Opteron Processor 252 (2.6 GHz). After all the docking jobs had finished, about 100 million poses were stored in Table "POSE", 80% of which were selected for ranking according to their interaction energies and efficiency. The computational time required for ranking a pose ranged from 1 s to 15 min depending on whether further calculations were needed for the pose and/or the convergence of quantum mechanical calculations.

The jobs were carried out by the Etna and Matterhorn clusters and some standalone computers simultaneously. The number of processors varied dynamically depending on the availabilities. The load of the database was low during the docking process using about 500 clients (<100 queries/s, about 100KB/s traffic of network, and <10% overall system load), because each docking client only communicated with the database 3 to 5 times per hour. Therefore, if the master of DVSDMS can distribute 60 molecules per second and a client requests 5 molecules per hour, the database server can support up to 43200 computing clients with nearly linear scalability if the clients are well synchronized. In contrast, during the ranking process, the load of the database was high (about 300 queries/s, 500−1000 KB/s traffic of network, and about 30% overall system load), because ranking 2000 poses usually took less than 5 min, and each of the 500 clients communicated with the database more than 12 times per hour. Note that docking and ranking could be combined sequentially to reduce the load of database.

CONCLUSIONS

DVSDMS uses freely available database software for efficient and automatic virtual screening distributed on Linux platforms. The essential concept of DVSDMS is the separation of data management from the main jobs in virtual screening, i.e., docking and ranking. In this way, the user has full flexibility on the choice of software for docking and ranking as well as hardware. Organized by DVSDMS, jobs are dispatched to each computing client to optimally exploit the available resources even in the case of heterogeneous hardware. Users not only can control and inspect the computing process but also attain consistent and logically organized data while computing is in progress or finished.

Because docking and ranking consist of many independent jobs, they are typically suited for a coarse-grained parallel architecture. In DVSDMS, computing clients do not communicate among each other but only with the database server. When a job is requested by a client, the database server scans the handling statuses and returns an unhandled job. In this way, a priori job partitioning is not required, and the overall computational load can be distributed equally to all computing clients. Moreover, the queue of jobs can be modified at any time; new jobs can be added to the computational pipeline by appending them to the database, while existing jobs can be removed before they start to run.

Upon performance tuning, the evaluation of the number of queries, the duration of queries, and the data flow indicate that the overall performance of DVSDMS is good. In particular, local cache and bulk updates reduce the query number; database index, proper storage engine, and database partitioning speed up queries; and data compression reduces

the data flow. Furthermore, since the database management system works as the master of DVSDMS, most of its sophisticated techniques (e.g., read write splitting and database cluster) can be applied directly to improve the performance of the master avoiding the overload without modifying the code for docking and ranking. In a docking benchmark, the master of DVSDMS built on a low cost Linux PC could distribute about 60 molecules per second. Furthermore nearly linear scalability of DVSDMS is expected up to 50000 nodes.

In the application presented here, docking the ZINC library into the receptor tyrosine kinase EphB4 with AutoDock and ranking poses under the control of DVSDMS, a low cost Linux PC was perfectly competent for the database server connected to about 500 computing clients. Since, the database management system MySQL and the program language Python are both open-source projects, DVSDMS can be applied in high-throughput virtual screening campaigns without restrictions typical of proprietary software.

AVAILABILITY OF DVSDMS

All scripts are available at http://biocroma.uzh.ch/zhou/dvsdms/.

ACKNOWLEDGMENT

We thank Armin Widmer for the continuous and extremely helpful support for the molecular modeling program Witnotp, and Philipp Schütz for useful suggestions and comments to the manuscript. This work was supported by a Swiss National Science Foundation grant to A.C. Most of the calculations were performed on the Etna and Matterhorn Beowulf cluster at the University of Zürich.

Supporting Information Available: Docking details, and the estimation of the maximum amount of molecules distributed per seconds for the MPI version of dock. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. Discovery of a Potent and Selective Protein Kinase CK2 Inhibitor by High-Throughput Docking. *J. Med. Chem.* **2003**, *46*, 2656–2662.
(2) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813–1818.
(3) Desai, P. V.; Patny, A.; Gut, J.; Rosenthal, P. J.; Tekwani, B.; Srivastava, A.; Avery, M. Identification of Novel Parasitic Cysteine Protease Inhibitors by Use of Virtual Screening. 2. The Available Chemical Directory. *J. Med. Chem.* **2006**, *49*, 1576–1584.
(4) Huang, D. Z.; Luthi, U.; Kolb, P.; Cecchini, M.; Barberis, A.; Caflisch, A. In Silico Discovery of Beta-Secretase Inhibitors. *J. Am. Chem. Soc.* **2006**, *128*, 5436–5443.
(5) Brown, S. P.; Muchmore, S. W. High-Throughput Calculation of Protein-Ligand Binding Affinities: Modification and Adaptation of the MM-PBSA Protocol to Enterprise Grid Computing. *J. Chem. Inf. Model.* **2006**, *46*, 999–1005.
(6) Kolb, P.; Huang, D.; Dey, F.; Caflisch, A. Discovery of Kinase Inhibitors by High-Throughput Docking and Scoring Based on a Transferable Linear Interaction Energy Model. *J. Med. Chem.* **2008**, *51*, 1179–1188.
(7) Kolb, P.; Kipouros, C. B.; Huang, D.; Caflisch, A. Structure-Based Tailoring of Compound Libraries for High-Throughput Screening: Discovery of Novel EphB4 Kinase Inhibitors. *Proteins* **2008**, *73*, 11–18.
(8) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
(9) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.
(10) Goodsell, D. S.; Olson, A. J. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins-Structure Function and Genetics* **1990**, *8*, 195–202.
(11) Zhang, S. X.; Kumar, K.; Jiang, X. H.; Wallqvist, A.; Reifman, J. DOVIS: An Implementation for High-Throughput Virtual Screening Using AutoDock. *Bmc Bioinformatics* **2008**, 9.
(12) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach To Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(13) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and Validation of a Modular, Extensible Docking Program: DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601–619.
(14) Peters, A.; Lundberg, M. E.; Lang, P. T.; Sosa, C. P. High Throughput Computing Validation for Drug Discovery Using the DOCK Program on a Massively Parallel System. *RedPaper* **2008**, *REDP-4410−00*.
(15) Mullen-Schultz, G. L.; Sosa, C. P. IBM System Blue Gene Solution. *Application Development.* **2007**, SG24−7179.
(16) Kolb, P.; Caflisch, A. Automatic and Efficient Decomposition of Two-Dimensional Structures of Small Molecules for Fragment-Based High-Throughput Docking. *J. Med. Chem.* **2006**, *49*, 7384–7392.
(17) Widmer, A., WITNOTP: A Computer Program for Molecular Modeling. Novartis: Basel, 1997.
(18) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
(19) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.
(20) Huang, D.; Caflisch, A. Efficient Evaluation of Binding Free Energy Using Continuum Electrostatics Solvation. *J. Med. Chem.* **2004**, *47*, 5791–5797.
(21) Zhou, T.; Huang, D.; Caflisch, A. Is Quantum Mechanics Necessary for Predicting Binding Free Energy. *J. Med. Chem.* **2008**, *51*, 4280–4288.
(22) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.
(23) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus Scoring for Ligand/protein Interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281–295.
(24) Gohlke, H.; Klebe, G. Statistical Potentials and Scoring Functions Applied to Protein-ligand Binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231–235.
(25) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(26) Momany, F. A.; Rone, R. Validation of the General-Purpose Quanta(R)3.2/Charmm(R) Force-Field. *J. Comput. Chem.* **1992**, *13*, 888–900.
(27) *SQLAlchemy*, 0.4.6; Michael Bayer: New York, NY, 2008.
(28) LaCour, J.; Haus, D.; Menten, d. G. Elixir. http://elixir.ematia.de/trac/ (accessed 31, Oct, 2008).
(29) MySQL::MySQL 5.1 Reference Manual::17.2.3 HASH Partitioning. http://dev.mysql.com/doc/refman/5.1/en/partitioning-hash.html (accessed 31, Oct, 2008).
(30) MySQL::MySQL 5.0 Reference Manual::13.1 The MyISAM Storage Engine. http://dev.mysql.com/doc/mysql/en/myisam-storage-engine.html (accessed 31, Oct, 2008).
(31) InnoDB Website. http://www.innodb.com/ (accessed 31, Oct, 2008).
(32) Gailly, J.; Adler, M. zlib Home Site. http://www.zlib.net/ (accessed 31, Oct, 2008).
(33) MySQL::MySQL 5.1 Reference Manual::10.5 Data Type Storage Requirements. http://dev.mysql.com/doc/refman/5.1/en/storage-requirements.html (accessed 31, Oct, 2008).
(34) ADT/AutoDockTools-AutoDock. http://autodock.scripps.edu/resources/adt/index_html (accessed 31, Oct, 2008).
(35) MySQL::The world's most popular open source database. http://www.mysql.com/ (accessed 31, Oct, 2008).

CI800295Q

Supporting Information

Data Management System for Distributed Virtual Screening

*Ting Zhou and Amedeo Caflisch**
Department of Biochemistry, University of Zürich,
Winterthurerstrasse 190, CH-8057
Zürich, Switzerland

DOCKING DETAILS

Before docking, the atom-specific affinity map files were created by AutoGrid.[1] The numbers of points in the x, y, and z directions were 62, 52, and 42, respectively, and the spacing between two adjacent grid points was 0.25 Å. The AutoDock[2] program was used to produce poses for further minimization and the custom ranking (Zhou et al., in preparation). To speed up the calculation, the maximum number of energy evaluations was set to 25000. The docking was followed by CHARMM[3] minimization using the CHARMm force field.[4] To suggest enough poses for minimization and testing of the ranking protocol, the hybrid genetic algorithm in AutoDock was run 400 times with different initial seeds. The poses were minimized with the rigid protein after docking, and the duplicated poses were eliminated by clustering using an all-atom RMSD cutoff of 0.01 Å.

THE ESTIMATION OF THE MAXIMUM AMOUNT OF MOLECULES DISTRIBUTED PER SECONDS FOR THE MPI VERSION OF DOCK

The estimation is based on "High Throughput Computing Validation for Drug Discovery Using the DOCK Program on A Massively Parallel System"[5] where the authors used a subset of 27005 drug-like ligands as a benchmark. Supposing that the efficiency of the "master" of MPI-DOCK is 100%, which is the upper limit, it took about 50000 seconds to finish docking on 256 Blue Gene/L processors. By increasing the number of processors to 16384, the efficiency diminished to about 55% due to the overload of the master. Therefore, the maximum amount of molecules distributed per seconds is about 19, which is calculated by $27005 \Big/ \left(\frac{50000 \times 256}{55\% \times 16384} \right)$.

REFERENCES

(1) AutoGrid-AutoDock. http://autodock.scripps.edu/wiki/AutoGrid (accessed 31, Oct, 2008).

(2) Goodsell, D. S.; Olson, A. J. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins-Structure Function and Genetics* **1990,** *8*, 195-202.

(3) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983,** *4*, 187-217.

(4) Momany, F. A.; Rone, R. Validation of the General-Purpose Quanta(R)3.2/Charmm(R) Force-Field. *J. Comput. Chem.* **1992,** *13*, 888-900.

(5) Peters, A.; Lundberg, M. E.; Lang, P. T.; Sosa, C. P. High Throughput Computing Validation for Drug Discovery Using the DOCK Program on a Massively Parallel System. *RedPaper* **2008,** *REDP-4410-00*.

Chapter 5

Complementing Ultrafast Shape Recognition with an Optical Isomerism Descriptor

Zhou, T.; Lafleur K.; Caflisch A. *Submitted*

Complementing Ultrafast Shape Recognition with an Optical Isomerism Descriptor

Ting Zhou, Karine Lafleur, and Amedeo Caflisch*

May 25, 2010

Department of Biochemistry, University of Zurich,
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
Phone: (+41 44) 635 55 21, Fax: (+41 44) 635 68 62
Email: caflisch@bioc.uzh.ch

Keywords: chirality descriptor, 3D database search, conformational isomerism, and conformer clustering

*To whom correspondence should be addressed.

Abstract: We introduce the mixed product of three vectors spanning four molecular locations as a descriptor of optical isomerism. This descriptor is very efficient as it does not require molecular superposition, and is very robust in discriminating between a given isomer and its mirror image. In particular, conformational isomers that are mirror images of each other, as well as optical isomers have opposite sign of the descriptor value. For efficient database searches, the optical isomerism descriptor can be used to complement an available ultrafast shape recognition (USR) method based solely on distances, which is not able to distinguish enantiomers. By an extensive comparison of the USR-based similarity score with an approach based on Gaussian molecular volume overlap, the accuracy and completeness of the former are discussed.

1 Introduction

Shape complementarity is essential in macromolecular recognition and binding of small molecules to proteins because of the sensitivity of the van der Waals energy at separations close to the optimal distance. There is abundant experimental evidence that small molecules with shape similar to known active compounds are likely to have similar biological activities.[1] Therefore, screening of databases of three-dimensional (3D) molecular structures can be performed by comparison of molecular shapes.[2–4] Several methods have been developed and applied in the past few decades to identify compounds similar to a query molecule.[5–10] They are useful whenever one or more inhibitors of a target protein are known particularly when the 3D structure of the protein is not available.

Recently, a method termed Ultrafast Shape Recognition (USR) has been developed for searching very large databases of molecular structures.[11] Despite its recent publication, USR has already been used in several drug design projects[3,12–15] because of its simplicity and efficiency. Importantly, the molecules do not need to be superposed. Only, the distances between each atom of the molecule and four molecular locations are calculated for USR: the molecular centroid

(ctd), the closest atom to ctd (cst), the farthest atom to ctd (fct), and the farthest atom to fct (ftf). The shape of a molecule is then encoded by 12 descriptors independent of the number of atoms. The first descriptor is the mean of atomic distances from ctd $\mu_1^{\text{ctd}} \equiv \frac{1}{N}\sum_{j=1}^{N} d_j^{\text{ctd}}$, where d_j is the distance of the jth atom from ctd, and N is the number of atoms in the molecule. The second descriptor is the square root of the second central moment of the distribution of the same atomic distances $\mu_2^{\text{ctd}} \equiv [\frac{1}{N}\sum_{j=1}^{N}(d_j^{\text{ctd}} - \mu_1^{\text{ctd}})^2]^{1/2}$. The third descriptor is the cubic root of the third central moment of the same distribution $\mu_3^{\text{ctd}} \equiv [\frac{1}{N}\sum_{j=1}^{N}(d_j^{\text{ctd}} - \mu_1^{\text{ctd}})^3]^{1/3}$ which is a measure of asymmetry. The remaining nine descriptors are calculated analogously using cst, fct, and ftf. Since only intramolecular distances are used in the 12 descriptors, the USR is not able to distinguish mirror images. Here, we supplement the original USR method[11] with an optical isomerism descriptor that is able to discriminate a molecule from its mirror image, and is therefore particularly useful for clustering conformers and searching 3D databases. Our extension of USR (called USR:OptIso) is first tested on three pairs of conformations of kinase inhibitors and 15 pairs of different types of isomers. Then similarity scores based on USR and USR:OptIso for 1.6×10^{10} pairs of conformers of 2.7 millions small molecules are compared with the ones based on Gaussian molecular volume overlap[16] calculated by ROCS (OpenEye Scientific Software).

2 Methods

Optical isomerism descriptor. Considerable efforts have been devoted to symmetry detection in chemistry and chemoinformatics. In particular, several methods have been developed to analyze chirality. These include two-dimensional descriptors[17–19] for the prediction of the major product of stereoselective reactions,[20–22] and three-dimensional descriptors (chiral topological indices) as complement to distance matrices in quantitative stereochemical structure-activity relationship models.[23–29]

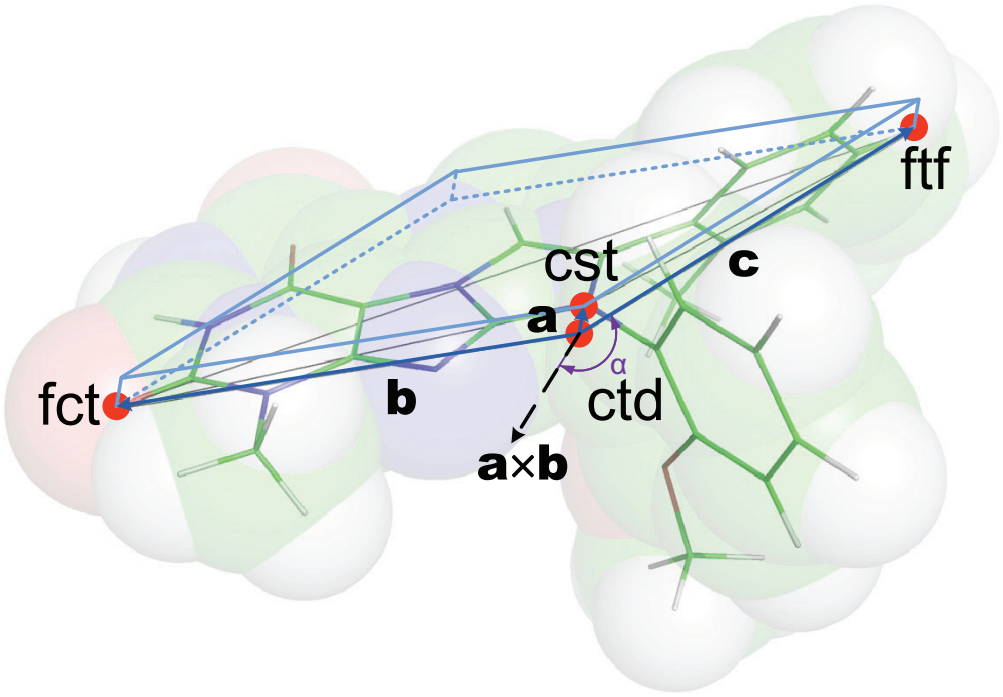
Here, the following vectors are introduced for the efficient evaluation of the optical isomerism

descriptor: $\mathbf{a} \equiv \mathbf{cst} - \mathbf{ctd}$, $\mathbf{b} \equiv \mathbf{fct} - \mathbf{ctd}$, and $\mathbf{c} \equiv \mathbf{ftf} - \mathbf{ctd}$, where **ctd**, **cst**, **fct**, and **ftf** are the vectors connecting the origin of the coordinates to each of the four molecular locations (Figure 1). The optical isomerism descriptor is defined as the cubic root of the scalar triple product (or mixed product) of **a**, **b**, and **c**, i.e., optical isomerism descriptor $\equiv [\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})]^{1/3}$. The cubic root is used to obtain a unit of length (Å) as for the other 12 descriptors. The ability of the optical isomerism descriptor to discriminate isomers is presented in Figure 2. For a molecule and its mirror image the 12 descriptors of the original USR method are identical because they only depend on distance distributions. However their optical isomerism descriptors are opposite because their four molecular locations are mirror images as well. In contrast, structural isomers, diastereoisomers, and other types of conformers that are not mirror image of each other have different distance distributions. Therefore, the first 12 descriptors are enough to discriminate them. The computational cost for evaluating the optical isomerism descriptor is neglectable, since the coordinates of the four molecular locations have to be calculated for the other 12 descriptors.

During the writing of this manuscript, Armstrong et al. reported a modification of USR that is able to distinguish enantiomers.[30] They use the *cross* product of two vectors spanning three of the four USR molecular locations (ctd, fct, and ftf) to define a fourth location which is different from cst. In contrast, the crucial component of our descriptor is the *triple* product of three vectors spanning all of the four locations. Moreover, Armstrong and collaborators replace three of the 12 USR descriptors (those involving cst) whereas we supplement the USR with the optical isomerism descriptor. Note that the atomic distances to ctd are different from that to cst as the separation between ctd and cst is usually between 0.4 and 2.0 Å for small molecules (Suppl. Mat. Figure S-1).

Similarity score. The inverse Manhattan distance is used as similarity score:[11]

$$S_{pq} = \frac{1\text{length unit}}{1\text{length unit} + \frac{1}{n}\sum_{i=1}^{n} |M_i^p - M_i^q|},$$



$$\text{optical isomerism descriptor} \equiv [\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})]^{1/3}$$

Figure 1: Optical isomerism descriptor. The four molecular locations of a conformer of compound **1** are denoted with red circles, and the three vectors **a**, **b**, and **c** with blue arrows. The optical isomerism descriptor is the cubic root of the volume of the parallelepiped with blue edges. The sign of the optical isomerism descriptor is negative for this conformer because **c** and $\mathbf{a} \times \mathbf{b}$ form an obtuse angle (violet angle α). The mirror image of this conformer has a positive value of the optical isomerism descriptor, and is shown in Figure 3.

Isomer1	Isomer2	Type of Isomerism	S_{pq}^{12}	S_{pq}^{13}	OID1	OID2	ROCS
		Structural isomerism: functional group isomerism	0.891	0.887	0.153	-0.032	0.999
		Structural isomerism: position isomerism	0.802	0.789	0.396	-0.106	0.825
		Structural isomerism: skeletal isomerism	0.826	0.833	0.838	0.911	0.892
		Conformational isomerism	1.000	0.889	0.815	-0.815	0.667
		Diastereoisomerism: cis/trans	0.979	0.972	0.456	0.334	0.685
		Diastereoisomerism: E/Z	0.922	0.928	0.000	0.000	0.302
		Chirality with one stereogenic centers: R/S	1.000	0.910	0.642	-0.642	0.997
		Chirality without stereogenic centers: allenes	1.000	0.884	0.849	-0.849	0.997
		Chirality without stereogenic centers: alkylidenecycloalkanes	1.000	0.860	1.054	-1.054	0.953
		Chirality without stereogenic centers: spiranes	1.000	0.987	-0.087	0.087	0.963
		Chirality without stereogenic centers: byphenyls - atroposomerism	1.000	0.964	-0.238	0.238	0.993
		Chirality without stereogenic centers: helicenes	1.000	0.790	1.725	-1.725	0.703
		Chirality without stereogenic centers - planar chirality: cyclophanes	1.000	0.759	2.053	-2.053	0.322
		Chirality without stereogenic centers - planar chirality: annulenes	1.000	0.863	1.035	-1.035	0.947
		Chirality without stereogenic centers - planar chirality: cyclalkenes	1.000	0.809	-1.514	1.514	0.957

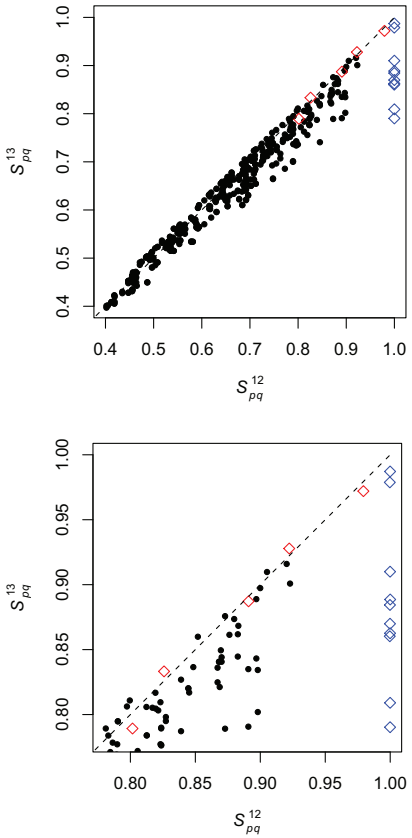


Figure 2: Application of the optical isomerism descriptors to isomers. All pairs of isomers can be distinguished by the USR:OptIso (13 descriptors) because the optical isomerism descriptors (OID1 and OID2) of mirror image isomers have opposite signs. The scatter plots show the pairwise USR comparisons of these 30 compounds (top right, full data set; bottom right, zoom-in on values close to 1). The original USR (S_{pq}^{12}) and USR:OptIso (S_{pq}^{13}) methods assign close similarities except for 10 pairs (blue diamonds) out of 435. The five pairs of isomers that can be distinguished by both S_{pq}^{12} and S_{pq}^{13} are denoted with red diamonds. The dashed line represents $S_{pq}^{13} = S_{pq}^{12}$. The last column of the table is the ROCS shape Tanimoto.[16]

where M_i^p is the ith descriptor of the conformation p. Note that all the USR descriptors have a unit of length, which is Å here, therefore the S_{pq} is dimensionless. The addition of 1 length unit at the denominator prevents a division by zero in the case of identical 3D structures, and yields a similarity score of 1 for them. The similarity scores with the optical isomerism descriptor (S_{pq}^{13}) and without (S_{pq}^{12}) are analyzed in the following.

First, it is interesting to compare the USR-based similarity scores with a metric based on superimposed volume. The correlation coefficient between the similarity score based on Gaussian molecular shape overlap and either S_{pq}^{12} or S_{pq}^{13} is 0.64 for the 30 compounds in Figure 2. This relatively low correlation is due to the fact that the similarity score evaluations are based on two different procedures. In the former, the similarity is calculated by volume-overlap percentage after structural superposition, whereas in USR, the similarity is evaluated using distributions of nuclei distances from the molecular centroids. Moreover, neither the iterative maximizing of the overlapped molecular volume in ROCS nor the maximum/minimum function for determining centroids in USR is continuous with respect to the Cartesian coordinates of the atomic nuclei, which are used as input. These two methods are compared extensively on around 1.6×10^{10} conformation pairs in the section Results and Discussion.

3 Results and Discussion

Mirror images, clustering, and database searches. A recently published inhibitor of the receptor tyrosine kinase Ephrin type-B receptor 4 (EphB4)[13] is used to illustrate the usefulness of the optical isomerism descriptor (Figure 3). The similarity score S_{pq}^{13} is able to distinguish the two mirror image conformers of compound **1** because of the opposite sign of their optical isomerism descriptors. In contrast, the two conformers of **1** have identical 12 descriptors based on the original USR method.[11]

The optical isomerism descriptor is useful for clustering as it can distinguish between different

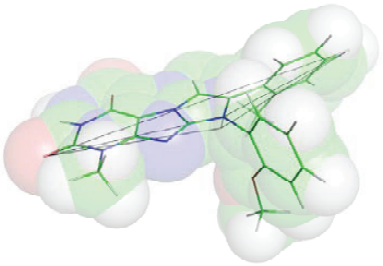
conformers/isomers that would be clustered together by the original USR. In our previous study,[13] multiple conformers of **1** were generated by systematic bond-rotation and optimized to their nearest local minima using density functional theory. Two local minima were then considered identical if their similarity score S_{pq} was higher than 0.999. Interestingly, the opposite sign of the optical isomerism descriptor contributes significantly to the identification of mirror images (or pairs of conformers very close to mirror images), in particular when S_{pq}^{12} is close to 1 (Figure 4). Furthermore, the USR:OptIso was tested on two pairs of isomers of recently published kinase inhibitors (**2** and **3** in Figure 5 and Figure 6, respectively).[31,32] The first 12 descriptors of USR have identical values, while the optical isomerism descriptor reduces the similarity score from 1 to 0.705 for compound **2** and from 1 to 0.650 for compound **3**.

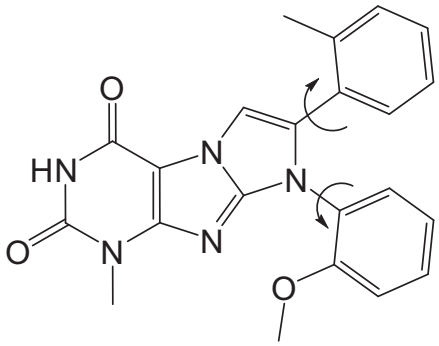
The optical isomerism descriptor can be used for searching (multi-)conformational libraries. As an example, using the similarity score S_{pq}^{13} yields only the conformer similar to the query whereas the conformations that are similar to its mirror image might be retrieved erroneously if one neglects the optical isomerism descriptor.

Finally, it is necessary to verify that similar conformers of a given molecule yield very similar values of S_{pq}^{12} and S_{pq}^{13}. A set of 100 similar structures of the protein kinase inhibitor PP2[33] was used for assessing the robustness upon minor structural change of the original USR and USR:OptIso. A scatter plot is presented in Suppl. Mat. Figure S-2. This test indicates that slight changes in the coordinates yield minor changes in both S_{pq}^{12} and S_{pq}^{13} when they are close to 1. Note that the high similarity range (i.e., values close to 1) is the most relevant case for virtual screening as only a small fraction of hits can be tested in practice.

Potential limitations of the optical isomerism descriptor. The optical isomerism descriptor is the cubic root of the (signed) volume of the parallelepiped defined by three vectors connecting four molecular locations. It is therefore equal to zero whenever the four molecular locations are coplanar. To estimate the frequency of the coplanarity of these four locations, we calculated the

Isomer *p* of compound **1**





$$\Downarrow$$

$$M^p = (3.93, 1.47, -0.83, 3.93, 1.50, -0.85, 7.23, 3.21, -2.39, 7.01, 3.10, -1.39, -2.64)$$

$$\Downarrow$$

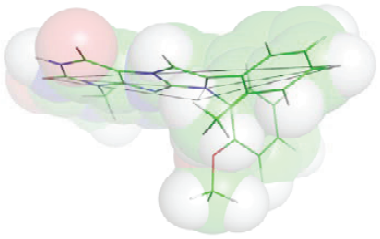
$$S_{pq} = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} |M_i^p - M_i^q|} \Rightarrow \begin{cases} n = 12 & S_{pq}^{12} = 1.000 \\ n = 13 & S_{pq}^{13} = 0.711 \end{cases}$$

$$\Uparrow$$

$$M^q = (3.93, 1.47, -0.83, 3.93, 1.50, -0.85, 7.23, 3.21, -2.39, 7.01, 3.10, -1.39, 2.64)$$

$$\Uparrow$$

Isomer *q* of compound **1**



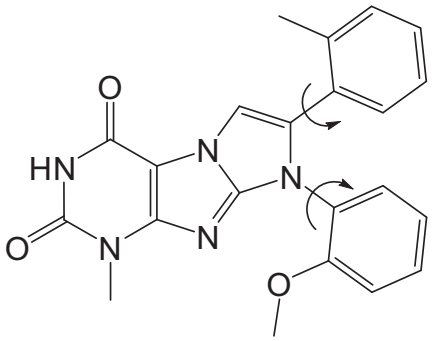


Figure 3: The two mirror images of **1** generated by systematic bond-rotation are distinguished by the optical isomerism descriptor. The four molecular locations are connected with black lines on the 3D structures which are shown with sticks and transparent CPK models. This example shows the high discriminating power of the optical isomerism descriptor whose usage results in a low similarity score $S_{pq}^{13} = 0.711$ for the two mirror images of **1** whereas they are not distinguished by the original USR method ($S_{pq}^{12} = 1.000$).

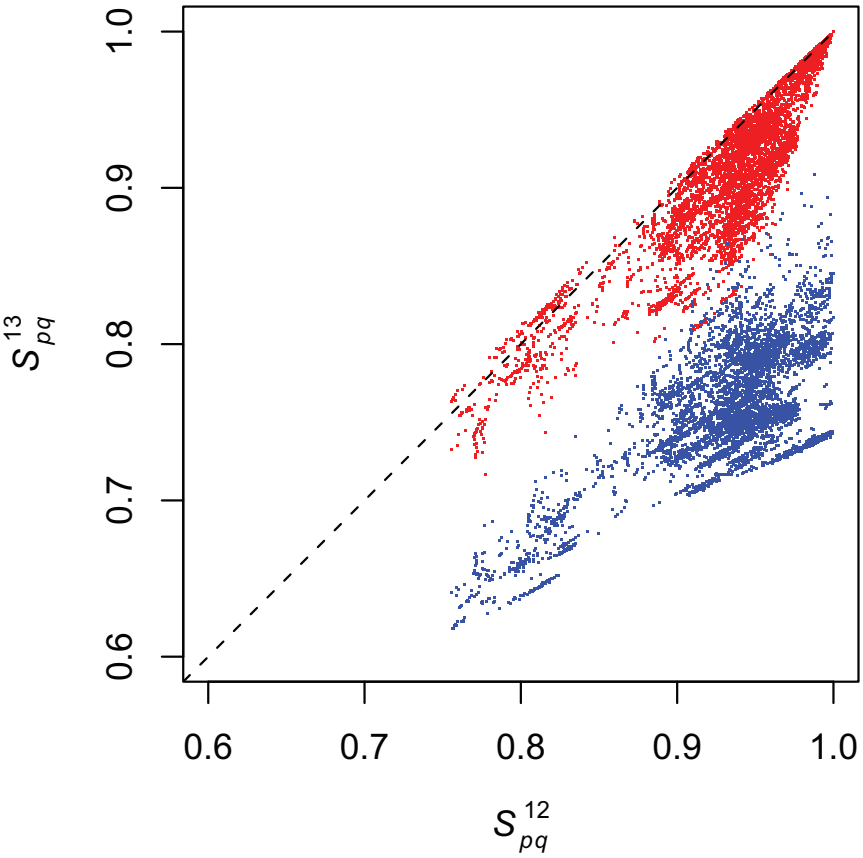


Figure 4: Scatter plot of the similarity scores with the optical isomerism descriptor (S_{pq}^{13}) and without (S_{pq}^{12}) for the 10,585 pairs of 146 local minima[13] of **1**. The color of each data point illustrates the sign of the optical isomerism descriptor (same and opposite signs are in red and blue, respectively). The dashed line represents $S_{pq}^{13} = S_{pq}^{12}$.

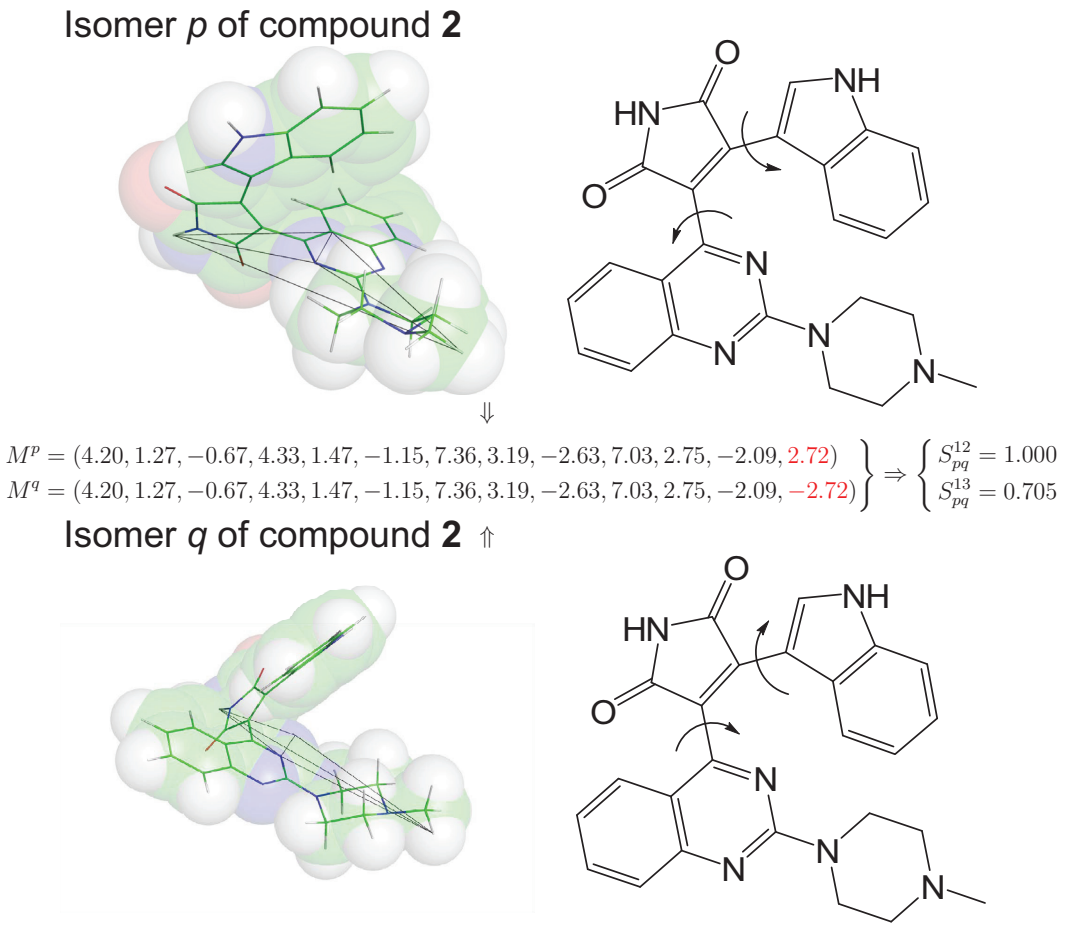


Figure 5: The two conformers of an inhibitor of protein kinase C are distinguished by the optical isomerism descriptor. The conformer of compound **2** observed in the X-ray structure[31] is shown in the top, while its mirror image is shown in the bottom. The similarity score decreases from 1 to 0.705 by taking into account the optical isomerism descriptor.

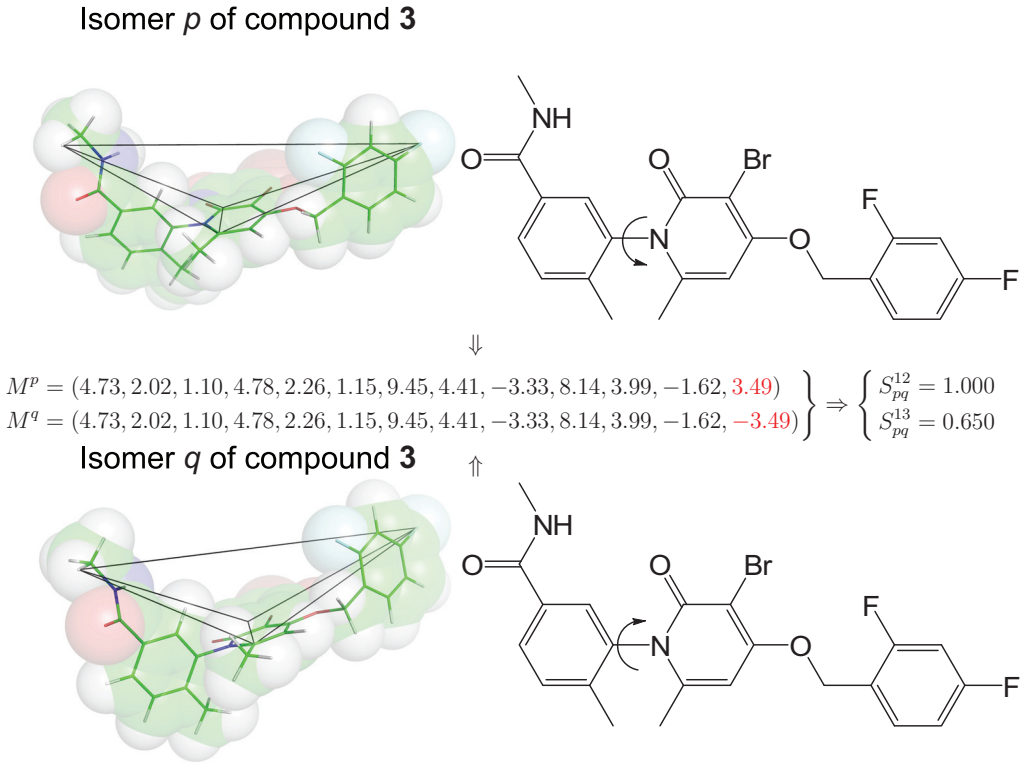


Figure 6: The two atropisomers of a p38 MAP kinase inhibitor are distinguished by the optical isomerism descriptor. The active conformer, which is >100-fold more potent than the other atropisomer,[32] is shown in the top, while its mirror image is shown in the bottom. The similarity score decreases from 1 to 0.650 by taking into account the optical isomerism descriptor.

optical isomerism descriptor of 100,812,356 poses of about 2.7 million molecules (downloaded from the 2007 version of the ZINC library[34]) generated by high-throughput docking into EphB4.[35] Strikingly, only 28 poses (of 24 molecules) have an optical isomerism descriptor smaller than 0.01 in absolute value, which indicates that the optical isomerism descriptor is able in the vast majority of cases to clearly discriminate a given isomer from its mirror image. Note that a value of the optical isomerism descriptor different from zero does not necessarily imply that a molecule is chiral. For this reason, we prefer to use the term "optical isomerism" rather than "chirality" descriptor.

Comparison between similarity scores calculated by USR and ROCS. To compare similarity scores calculated by USR and USR:OptIso with ROCS shape Tanimoto scores (the maximum ratio of the overlap of a pair of Gaussian molecular volumes) the aforementioned 100 million poses of about 2.7 million molecules from ZINC were used to generate 1.6×10^{10} pairs of conformations. Notably, these two methods have very different algorithms to evaluate molecular shape similarity. Cartesian coordinates are the only input required to calculate USR scores, whereas ROCS also needs atomic radii to evaluate the Gaussian molecular volume. To compare the conformation pairs filtered by different similarity cutoffs, the "accuracy" was defined as the number of pairs for which both scores exceeded the cutoff divided by the pairs for which only the USR similarity score exceeded the cutoff (i.e., C/(A+C) in Figure 7). In the same way, the "completeness" was defined as the ratio of the number of pairs for which both scores exceeded the cutoff to the pairs for which only ROCS shape Tanimotos exceeded the cutoff (i.e, C/(B+C) in Figure 7).

Both accuracy and completeness increase monotonously with the similarity cutoff (Figure 7). The accuracy of USR:OptIso has improved compared with the original USR, because the mirror images of the query conformation have been eliminated by the opposite optical isomerism descriptors. For instance, 91.25% of the conformation pairs that have USR:OptIso scores ≥ 0.968 also have ROCS shape Tanimotos ≥ 0.968, whereas the percentage decreases to 57.17% for the

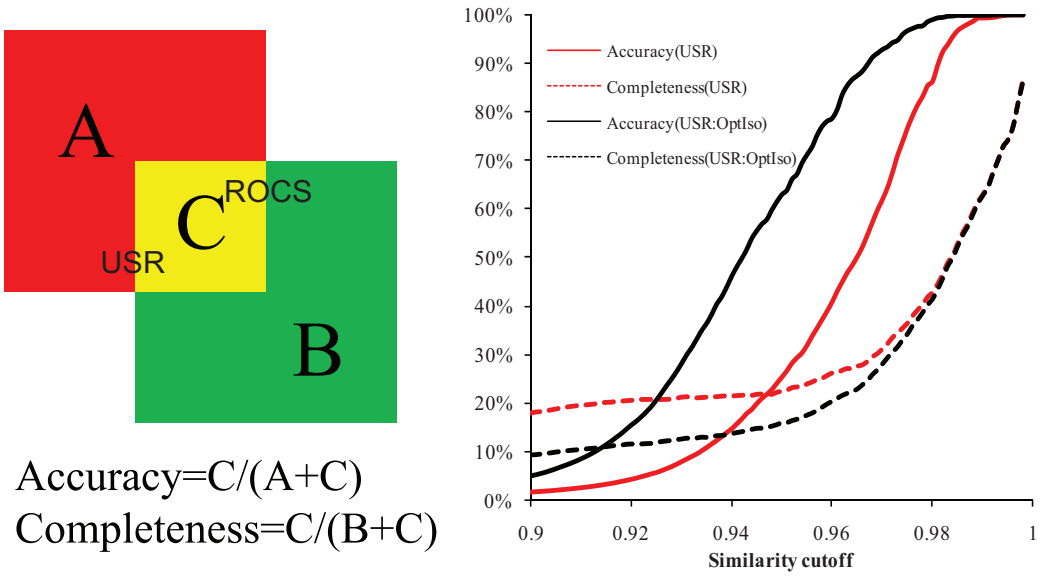


Figure 7: Accuracy and completeness of USR similarity score in reproducing ROCS shape Tanimoto. In the left figure, the green and the red squares are the sets of conformation pairs for which USR similarity scores and ROCS shape Tanimoto satisfy a particular cutoff, respectively. Their overlap (yellow part) represents the conformation pairs for which both similarity scores satisfy the cutoff. In the right figure, the cutoff varies from 0.9 to 1.0, which covers the range interesting for virtual screening applications. The solid and dashed lines represent accuracy and completeness, respectively, while the red and black color denote the USR and USR:OptIso results, respectively.

original USR. If the similarity score cutoff is set to 0.98, the accuracy of the USR:OptIso increases to 99.02% compared to 86.30% for the original one. Figure 8 (a) shows "the most inaccurate" example whose USR score is higher than the ROCS shape Tanimoto.

The completeness of USR methods is low. One of the main reasons is that USR overestimates the difference of conformation pairs that are different at the extremity of the molecule where fct and/or ftf are defined. Figure 8 (b) shows an example of a conformation pair whose ROCS shape Tanimoto is higher than the USR similarity score. In this example, USR overestimates the conformational difference because of a change of the ftf location. The low completeness is not surprising as ROCS shape Tanimoto distinguishes atomic elements by different van der Waals radii whereas both USR and USR:OptIso treat all atoms equally. Moreover, the optimal volume overlap has to be calculated in ROCS shape Tanimoto for every pair of conformations while the USR methods are significantly more efficient as they use only interatomic distances.

4 Conclusions

The optical isomerism descriptor (defined as the mixed product of the three vectors from the molecular centroid to the three molecular locations cst, fct, ftf) is an efficient and robust tool for shape comparison. It can be used as a supplement of the original 12 USR descriptors, which are based solely on distance distributions, while the optical isomerism descriptor is able to distinguish mirror images. It is therefore helpful for analyzing molecules with stereogenic centers, atropisomerism, and in the clustering of conformers generated by systematic bond-rotation. Moreover, it can be used for the efficient search of molecular conformations that are superposable on the query structure. Finally, a comparison of the USR similarity score with the ROCS shape Tanimoto shows that both accuracy and completeness increase monotonously with the similarity score cutoff. The accuracy of the USR:OptIso similarity score is always higher than the one based on the original USR, and the completeness of USR:OptIso is close to the one of USR in high

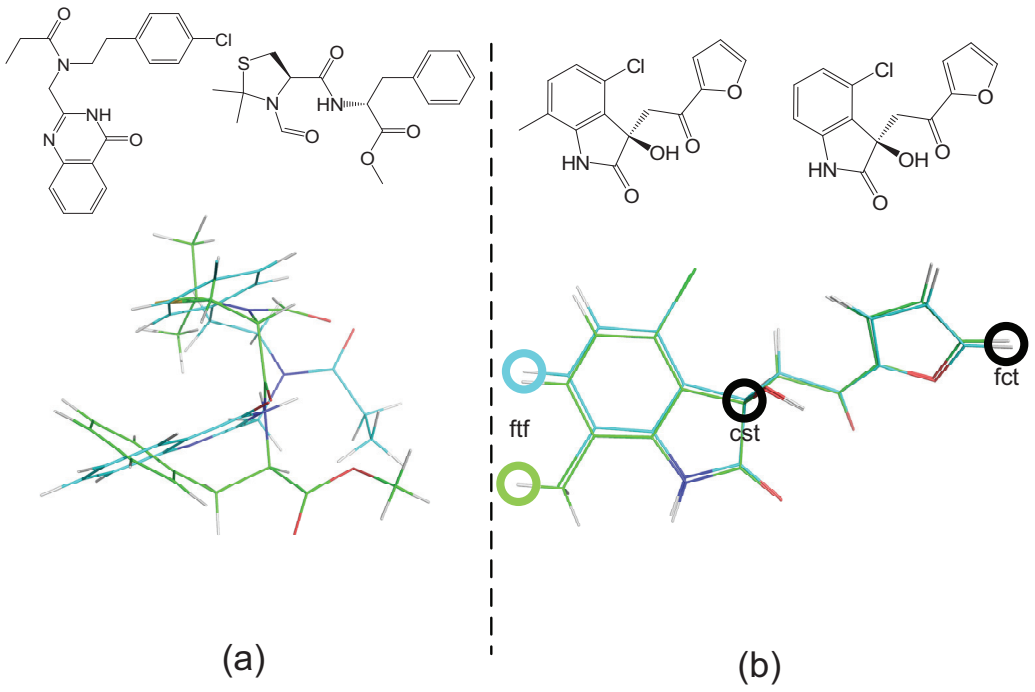


Figure 8: Examples of large discrepancies between USR and ROCS shape Tanimoto. (a) High USR similarity score but low ROCS shape Tanimoto. The overlaid conformations were optimized by ROCS. The USR and USR:OptIso similarity scores are 0.9807 and 0.9810. The ROCS shape Tanimoto is 0.676. This is the only case out of 1.6×10^{10} conformation pairs whose USR is larger than 0.98 and ROCS shape Tanimoto smaller than 0.7. (b) Low USR similarity score but high ROCS shape Tanimoto. The USR and USR:OptIso similarity scores are 0.6988 and 0.6981. The ROCS shape Tanimoto is 0.999. These two molecules are different in only one substituent. The additional methyl group in the green conformation changes the location of ftf from the phenyl hydrogen (of the cyan conformation) to the methyl hydrogen.

similarity ranges, which are relevant for virtual screening.

Acknowledgement. We thank Dr. Ballester for providing a detailed description of USR. We thank OpenEye Scientific Software for providing an academic license of ROCS.

Source Code. The source code for calculating the optical isomerism descriptor is available at http://code.google.com/p/usrchirality/.

References

1. McGaughey, G.; Sheridan, R.; Bayly, C.; Culberson, J.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.; Cornell, W. *J Chem Inf Model* **2007**, *47*, 1504–1519.
2. Rush, T.; Grant, J.; Mosyak, L.; Nicholls, A. *J Med Chem* **2005**, *48*, 1489–1495.
3. Ballester, P. J.; Finn, P. W.; Richards, W. G. *J Mol Graph Model* **2009**, *27*, 836–845.
4. Kinnings, S.; Jackson, R. *J Chem Inf Model* **2009**, *49*, 2056–2066.
5. Artymiuk, P.; Bath, P.; Grindley, H.; Pepperrell, C.; Poirrette, A.; Rice, D.; Thorner, D.; Wild, D.; Willett, P.; Allen, F. *J Chem Inf Comput Sci* **1992**, *32*, 617–630.
6. Sheridan, R.; Kearsley, S. *Drug Discov Today* **2002**, *7*, 903–911.
7. Ruiz, I. L.; García, G. C.; Gómez-Nieto, M. A. *J Chem Inf Model* **2005**, *45*, 1178–1194.
8. Haigh, J.; Pickup, B.; Grant, J.; Nicholls, A. *J Chem Inf Model* **2005**, *45*, 673–684.
9. Hawkins, P.; Skillman, A.; Nicholls, A. *J Med Chem* **2007**, *50*, 74–82.
10. Cao, Y.; Jiang, T.; Girke, T. *Bioinformatics* **2008**, *24*, i366–i374.
11. Ballester, P. J.; Richards, W. G. *J Comput Chem* **2007**, *28*, 1711–1723.

12. Cannon, E.; Nigsch, F.; Mitchell, J. *Chem Cent J* **2008**, *2*, 3.

13. Lafleur, K.; Huang, D.; Zhou, T.; Caflisch, A.; Nevado, C. *J Med Chem* **2009**, *52*, 6433–6446.

14. Li, H.; Huang, J.; Chen, L.; Liu, X.; Chen, T.; Zhu, J.; Lu, W.; Shen, X.; Li, J.; Hilgenfeld, R.; Jiang, H. *J Med Chem* **2009**, *52*, 4936–4940.

15. Ballester, P. J.; Westwood, I.; Laurieri, N.; Sim, E.; Richards, W. G. *J R Soc Interface* **2010**, *7*, 335–342.

16. Grant, J. A.; Gallardo, M. A.; Pickup, B. T. *J Comput Chem* **1996**, *17*, 1653–1666.

17. Randic, M.; Razinger, M. *J Chem Inf Comput Sci* **1996**, *36*, 429–441.

18. Randic, M. *J Chem Inf Comput Sci* **2001**, *41*, 639–649.

19. Zhang, Q. Y.; de Sousa, J. A. *J Chem Inf Model* **2006**, *46*, 2278–2287.

20. de Sousa, J. A.; Gasteiger, J. *J Chem Inf Comput Sci* **2001**, *41*, 369–375.

21. de Sousa, J. A.; Gasteiger, J. *J Mol Graph Model* **2002**, *20*, 373–388.

22. de Sousa, J. A.; Gasteiger, J.; Gutman, I.; Vidović, D. *J Chem Inf Comput Sci* **2004**, *44*, 831–836.

23. Schultz, H. P.; Schultz, E. B.; Schultz, T. P. *J Chem Inf Comput Sci* **1995**, *35*, 864–870.

24. Golbraikh, A.; Bonchev, D.; Tropsha, A. *J Chem Inf Comput Sci* **2001**, *41*, 147–158.

25. Capozziello, S.; Latranzi, A. *Chirality* **2003**, *15*, 227–230.

26. Fujita, S. *Theor Chem Acc* **2005**, *113*, 80–86.

27. Yang, C. S.; Zhong, C. L. *QSAR Comb Sci* **2005**, *24*, 1047–1055.

28. Dervarics, M.; Otvos, F.; Martinek, T. A. *J Chem Inf Model* **2006**, *46*, 1431–1438.

29. Natarajan, R.; Basak, S.; Neumann, T. *J Chem Inf Model* **2007**, *47*, 771–775.

30. Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Richards, W. G. *J Mol Graph Model* **2009**, *28*, 368–370.

31. Wagner, J.; Matt, P. v.; Sedrani, R.; Albert, R.; Cooke, N.; Ehrhardt, C.; Geiser, M.; Rummel, G.; Stark, W.; Strauss, A.; Cowan-Jacob, S.; Beerli, C.; Weckbecker, G.; Evenou, J.-P.; Zenke, G.; Cottens, S. *J Med Chem* **2009**, *52*, 6193–6196.

32. Xing, L.; Shieh, H.; Selness, S.; Devraj, R.; Walker, J.; Devadas, B.; Hope, H.; Compton, R.; Schindler, J.; Hirsch, J.; Benson, A.; Kurumbail, R.; Stegeman, R.; Williams, J.; Broadus, R.; Walden, Z.; Monahan, J. *Biochemistry (Mosc)* **2009**, *48*, 6402–6411.

33. Sturz, A.; Bader, B.; Thierauch, K.; Glienke, J. *Biochem Biophys Res Commun* **2004**, *313*, 80–88.

34. Irwin, J.; Shoichet, B. *J Chem Inf Model* **2005**, *45*, 177–182.

35. Zhou, T.; Caflisch, A. *J Chem Inf Model* **2009**, *49*, 145–152.

Supporting Information

Complementing Ultrafast Shape Recognition with an Optical Isomerism Descriptor

*Ting Zhou, Karine Lafleur and Amedeo Caflisch**
Department of Biochemistry, University of Zürich,
Winterthurerstrasse 190, CH-8057
Zürich, Switzerland

Complementing Ultrafast Shape Recognition with an Optical Isomerism Descriptor

Ting Zhou, Karine Lafleur, and Amedeo Caflisch*

May 25, 2010

Department of Biochemistry, University of Zurich,

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Phone: (+41 44) 635 55 21, Fax: (+41 44) 635 68 62

Email: caflisch@bioc.uzh.ch

Keywords: chirality descriptor, 3D database search, conformational isomerism, and conformer clustering

*To whom correspondence should be addressed.

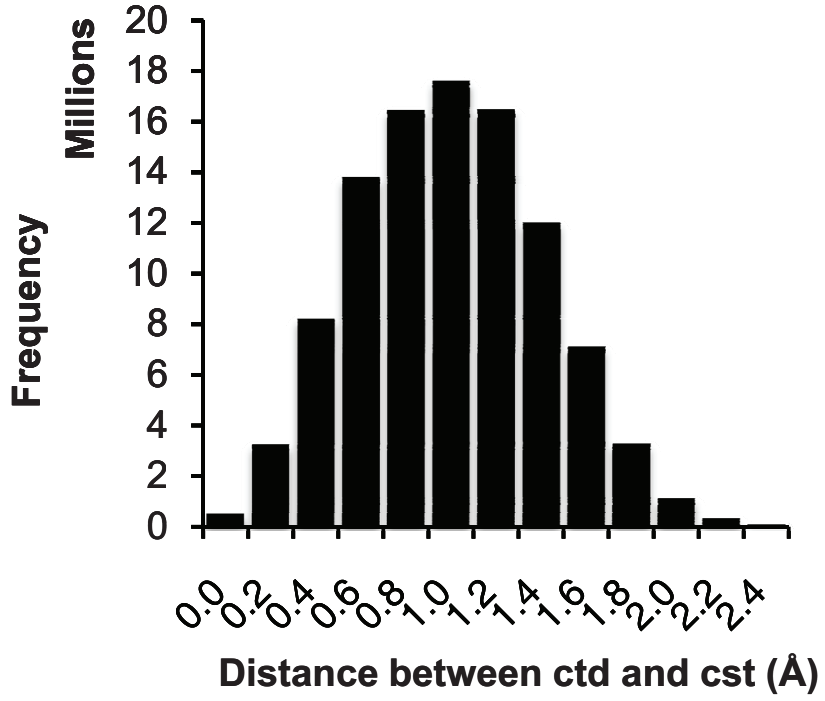


Figure S-1: The distribution of ctd–cst distances of 100 million poses of 2.7 million of molecules. These molecules are from a subset of the ZINC library with molecular weight less than 500 Da.

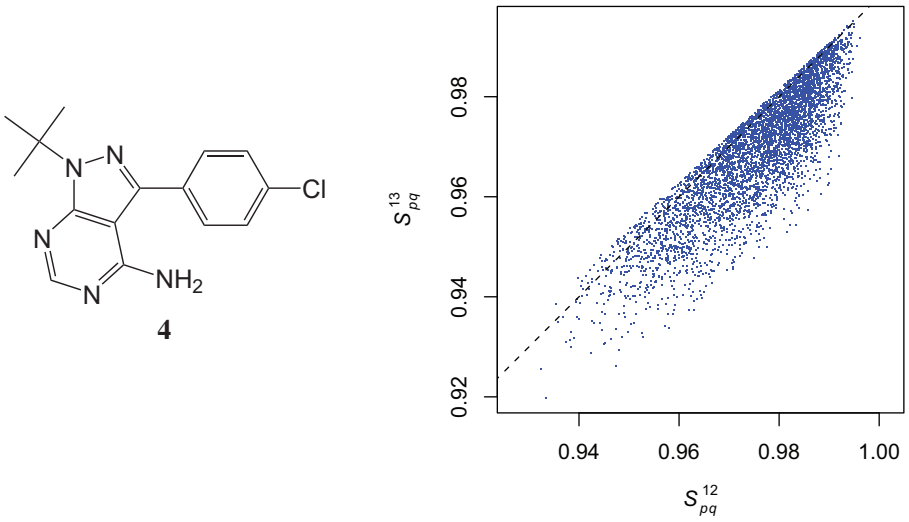


Figure S-2: Pairwise comparison of 100 little fluctuated conformers of compound **4**. The dashed line represents $S_{pq}^{13} = S_{pq}^{12}$.

Conclusions and Outlook

Several quantum mechanical(QM) methods already play an important role in many phases of computer-aided drug discovery (CADD), and will have a stronger impact in the future because of the ever growing computing power and development of efficient algorithms. The work in this thesis focuses on the use of QM to improve methods for calculating the binding free energy, and the implementation of the QM method in high-throughout virtual screening. The compromise between accuracy and efficiency is a perpetual issue in the CADD applications especially for QM methods. One of the main conclusions of this thesis is that it is important to select the most appropriate technique at each phase of drug development, and QM methods should be selected only if there is a real advantage with respect to the classical approaches. The initial phase of CADD, e.g., high-throughput docking, which is useful to identify hit compounds, requires full sampling of conformations of the small molecules within the protein binding site. Such extensive sampling calls for approximated energy functions and predicted properties thereof, which are usually calculated by classical force field methods or fast semi-empirical QM methods, e.g., QM probe methods. In the subsequent phase, hits have to be optimized to leads which does not require extensive sampling but high accuracy because of the small differences in the binding free energy. Therefore QM methods should be applied on the hits to shed light on the energetics of binding. The QM methods are particularly important to capture charge transfer and polarization effects, which are usually pronounced in systems containing metal atoms or charged groups, and/or dispersion forces which play a significant role in the interactions of conjugated π systems. Importantly, before starting CADD it is necessary to evaluate the status of the project, which in turn dictates the number and diversity of molecules to be evaluated and the demand of accuracy, and to select the most appropriate approach accordingly.

However, the CADD methods with static descriptions of molecules (like those described in this thesis) seem to reach a plateau-like region instead of the globally optimum after years of efforts. The limited computing power and an insufficient description of dynamics of both the target and ligands hinder the analysis of dynamic intra/intermolecular interactions under realistic high-throughput circumstances. Consequently, the molecular fluctuations, which contribute to the entropic part of binding free energy, and the induced-fit, which contributes to the enthalpic part, cannot be evaluated adequately. In fact, based on rigorous classical statistical mechanics, molecular dynamics (MD) or Monte Carlo (MC) was extended for calculating free energy differences in the late 1970s either by free energy perturbation or by thermodynamic integration. The problem of integration of insufficient sampling of MD/MC can be partially resolved by limiting the scale of virtual screening and using more computing power in calculation, However, inaccuracy and inadequacy of the force field, particularly for small molecules and metal ions, quickly emerge when these calculations are performed nowadays. The weakness of parameterization of the force field can be naturally overcome

with QM methods, but QM methods still have difficulties in yielding accurate weak London dispersion forces with acceptable amount of calculation for dynamics of macromolecules.

The CADD is an interdisciplinary research of chemistry, biology, mathematics and computer science. This intrinsic character calls for importation of concepts from other fields of science, not only the newfangled technology but also the classical methodology. In the other part of this thesis, a professional database management system developed by computer scientists is applied in high-throughput virtual screening. The open-source programs MySQL is used for handling large amount of data of molecular properties and calculated energy terms efficiently with small amount of extra programming. Well-organized data is required for the further systematic research to reveal the relationship among targets and ligands by taking advantage of chemogenomic and bioinformatic network analysis, which, as a return, will benefit the computer-aided screening and design of compounds with a desired activity and property profile.

Acknowledgements

I heartily thank Prof. Amedeo Caflisch for the greatest help in all aspects of my research. He not only taught me how to research but also affected me with his accurate and rigorous scientific attitude. I also thank Prof. Kim Baldridge for being part of my PhD committee.

I would like to acknowledge Dr. Danzhi Huang, who guided me through my first year patiently. I am thankful that Dr. Fabian Dey, Dr. Philipp Schütz, Dr. Peter Kolb, Dr. Ran Friedman, Dr. Riccardo Pellarin, François Marchand, Dr. Stefanie Muff, and Dr. Dariusz Ekonomiuk offer me wonderful computer hardware and software support. I thank Prof. Cristina Nevado, Dr. Danzhi Huang and Karine Lafleur for cooperations in my research and writing manuscripts. I appreciate Sandra Rennebaum for the efficient assist of German translation of the abstract.

I want to offer my regards to all the group members who support me in any respect during my PhD research, and all the colleagues in biochemistry department of University of Zurich for the cooperative environment.

Special thanks to my parents Linquan Zhou and Naining Yang for encouraging me to study in Switzerland, and my wife Yu Chen for taking care of the family.

Curriculum Vitae

Personal Information

Surname:	Zhou
First name:	Ting
Current address:	Wallisellenstrasse 473, 8050 Zürich, Switzerland
Nationality:	Chinese
Birthday:	09.08.1981
Birthplace:	Nanjing (China)

Education

University of Zurich, Zurich, Switzerland 08/2006 – Present
Ph.D. Program in Computational Structural Biology
Concentration in new algorithms and tools for computer-aided drug design
2009 Chinese Government Award for Outstanding Self-financed Students Abroad

Nanjing University, Nanjing, China 09/2003 – 06/2006
Master of Science in Chemistry
Concentration in theory and computational simulation of microfluidics and electrochemistry

Nanjing University, Nanjing, China. 09/1999 – 06/2003
Bachelor of Science in Department for Intensive Instruction
Major in Chemistry, Biology, and Physics
People's Scholarship of Nanjing University

Nanjing No. 7 Middle School, Nanjing, China. 09/1993 – 06/1999

Peer-reviewed Publications

- **Zhou, T.**; Caflisch A., High-throughput Virtual Screening using Quantum Mechanical Probes: Discovery of Selective Kinase Inhibitors, *ChemMedChem* in press.
- **Zhou, T.**; Lafleur, K.; Caflisch A., Complementing Ultrafast Shape Recognition with an Optical Isomerism Descriptor, *submitted*.
- Huang, D.; **Zhou, T.**; Lafleur, K.; Nevado, C.; Caflisch A., Kinase selectivity potential for ATP-competitive inhibitors: A network analysis. *Bioinformatics* 2010, 26(2), 198–204.
- Lafleur, K.; Huang, D.; **Zhou, T.**; Nevado, C.; Caflisch A., Structure-Based Optimization of Potent and Selective Inhibitors of the Tyrosine Kinase EphB4. *Journal of Medicinal Chemistry* 2009, 52 (20), 6433–6446.
- **Zhou, T.**; Huang, D.; Caflisch A., Quantum Mechanics in Structure-based Drug Design. *Current Topics in Medicinal Chemistry* 2010, 10 (1), 33–45.
- Bodenreider, C.; Beer, D.; Keller, T. H.; Sonntag, S.; Wen, D.; Yap, L.; Yau, Y. H.; Shochat, S. G.; Huang, D.; **Zhou, T.**; Caflisch, A.; Su, X.; Ozawa, K.; Otting, G.; Vasudevan, S. G.; Lescar, J.; Lim, S. P., A fluorescence quenching assay to discriminate between specific and non-specific inhibitors of dengue virus protease. *Analytical Biochemistry* 2009, 395 (2), 195–204.
- **Zhou, T.**; Caflisch A., Data management system for distributed virtual screening. *Journal of Chemical Information and Modeling* 2009, 49 (1), 145–152.
- **Zhou, T.**; Huang, D.; Caflisch A., Is Quantum Mechanics Necessary for Predicting Binding Free Energy? *Journal of Medicinal Chemistry* 2008, 51 (14), 4280–4288.